Ph.D. DISSERTATION

# Visual Analysis for MicroRNA and mRNA Expression Profile Data

# 마이크로 RNA 와 mRNA 표현형 데이터를 위한 시각적 분석

AUGUST 2016

DEPARTMENT OF ELECTRICAL ENGINEERING &
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Daekyoung Jung

# Visual Analysis for MicroRNA and mRNA Expression Profile Data

마이크로 RNA 와 mRNA
표현형 데이터를 위한 시각적 분석

지도교수 서 진 욱

이 논문을 공학박사 학위논문으로 제출함

2016 년 4 월

서울대학교 대학원

전기 · 컴퓨터 공학부

정 대 경

정대경의 공학박사 학위논문을 인준함

2016 년 6 월

위 원 장 : 김 선 (인)
부위원장 : 서 진 욱 (인)
위 원 : 장 병 탁 (인)
위 원 : 김 보 형 (인)
위 원 : 이 정 진 (인)

# Abstract

## Visual Analysis for MicroRNA and mRNA Expression Profile Data

Daekyoung Jung

Department of Electrical Engineering & Computer Science

College of Engineering

The Graduate School

Seoul National University

MicroRNAs (miRNA) are short nucleotides that down-regulate its target genes. Various miRNA target prediction algorithms have used sequence

complementarity between miRNA and its targets. Recently, other algorithms tried to improve sequence based miRNA target prediction by exploiting miRNA-mRNA expression profile data.

Some web-based tools are also introduced to help researchers predict miRNAs targets from miRNA-mRNA expression profile data; However, there is still a demand for a miRNA-mRNA visual analysis tool that include quality miRNA prediction algorithms and more interactive visualizations.

We presented two techniques for miRNA-mRNA interaction visualizations, Bipartite Treemap and enhanced node-link diagram. Bipartite Treemap is a new visualization technique for miRNA-mRNA interaction network that resolves occlusion problem. Enhanced node-link diagram provides interaction techniques that help users to explore miRNA-mRNA interaction network easily.

We designed and implemented miRTarVis, which is an interactive visual analysis tool that predicts miRNA targets by integrating sequence based and miRNA-mRNA expression profile based miRNA target

prediction algorithms, and visualizes the resulting miRNA-mRNA interaction network. miRTarVis has intuitive interface design in accordance with the analysis procedure of load, filter, predict, and visualize. It predicts miRNA targets by adopting Bayesian inference and MINE analyses, as well as conventional correlation and mutual information analyses. It visualizes a resulting miRNA-mRNA network in an interactive Bipartite Treemap as well as ehanced node-link diagram.

Using miRTarVis, we analyzed miRNA-mRNA expression profile data from an experiment over asthmatic and non-asthmatic fibroblasts exposed to obese visceral exosomes. In addition, we applied miRTarVis to miRNA-mRNA expression profile data from breast cancer cell lines data to show its efficacy. miRTarVis verified its efficacy by helping its users execute miRNA target prediction easily and gain insights from miRNA-mRNA expression profile data by its interactive visualization.

**keywords : MicroRNA, mRNA, Visualization, Gene Expression, MicroRNA target prediction, Visual analysis, Bipartite Treemap, miRTarVis**

**student number : 2010-20886**

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background and Motivation

Recent rapid advance in methodologies for measuring genomic and epigenomic data enables researchers to acquire whole-genome wide sequencing information with higher throughput than ever. These set of advanced technologies are called next generation sequencing (NGS) or high-throughput sequencing, as their high throughput dramatically overcomes previous sequencing technologies' limitations.

Also, NGS technologies supply more precise genetic data than before. For example, RNA sequencing (RNA-Seq) technique, a major application of NGS for transcriptional studies [1], can measure whole-genome wide gene expression with higher accuracy previous microarray technique

does [2]. Another example of NGS application is chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq), which can identify epigenetic information such as transcription factor binding sites or histone modification with higher specificity than previous ChIP-Chip method does [3].

One challenge of NGS is its large data size. Large data is being yielded by NGS technologies with reasonable low cost and short time, and it is expected that sequencing cost and speed will decrease [4], so it is expected that massive size of data will be generated. A study that conducts comparative genomic analysis between identical twins using NGS techniques [5] shows an example for data size for NGS.

There were billions of reads for whole-genome sequencing and ten millions of reads for transcriptomes and methylomes for each individual. As average length of reads is approximately 100 bp, and one byte of storage is required per base pair, the necessary storage size for saving experiment result is more than hundreds gigabytes. According to Zhang et al. [6], on average, NGS experiment generally generates terabytes of raw data. Therefore, a powerful computing system with large storage,

memories and many computing cores is necessary for storage, analysis, and visualization of NGS data [4] [7] [8].

As the data size from genomic and epigenomic experiments increases, the importance of the bioinformatic analysis tools is increasing, since bioinformatic analysis tools are essential to achieve full benefit of NGS data [7] [9]. The analysis of NGS data consists of data analysis techniques at manifold levels. Series of analysis techniques that process from raw sequencing data to end-user-understandable meta data is called pipeline.

The analysis techniques relatively close to the raw data are called up-stream analyses, and those relatively close to the user levels are called down-stream analyses. For example, a pipeline for analyzing RNA-Seq experiments normally consists of, from up-stream to down-stream, quality control and filtering, sequence alignment to the reference genome, transformation and normalization of read counts, identification for differentially expressed transcriptomes, and visualization of gene expression for end users. In this paper, we focus on

the down-stream analysis and visualization of microRNA-mRNA expression profile data.

NGS technologies enables researcher measure various important genomic factors in gene regulation, and microRNA (miRNA) is one of them. MiRNA is one of important gene expression regulators. Multicellular organisms regulate of gene expression to drive genetically homogeneous cells function heterogeneously [10]. Increasingly, it is recognized that regulatory RNAs (e.g. microRNAs, PIWI-interaction RNAs) play key roles in epigenetic processes that control differentiation and development [11]. Thus, understanding the regulation of the gene expression requires the integrated analysis of mRNA with regulatory RNAs. Among the regulatory RNAs miRNAs are relatively well characterized and studied.

MiRNAs are short highly processed oligonucleotides (approximately 22 nt) that carry out post-transcriptional regulation of target mRNAs by either degradation of the target mRNA or inhibition of protein translation [12] [13]. MiRNA researchers' main concern is to find miRNAs' targets since identification of miRNAs' targets can help

recognizing their biological functions. The methods for finding miRNAs' target can be categorized into three groups: experimental validation, sequence-based prediction, and miRNA-mRNA expression profile based prediction [14].

There are diverse experimental methods for validation of MiRNA's targets. Widely-used techniques include 1) measurement of transcriptome or proteome expression under transfection of mimic miRNAs or miRNA inhibitors, 2) attachment of reporters or labels to miRNAs and 3'-UTR of mRNAs, or 3) exploiting immunoprecipitation of RISC complexes [15] [16] [17].

However, experimental methods are expensive in terms of cost and time, miRNA target prediction algorithms are introduced. There are two types of miRNA target prediction algorithm. First type is sequence based prediction algorithms. They predict miRNA targets based on experimentally determined rules. The rules take into account 1) sequence alignment between seed region (nucleotides 2-7) of miRNA and 3'-UTR of mRNA, 2) possible target sites in coding region or 5' UTR of

mRNA, 3) evolutionary conservation of the target mRNA sequence among related species, or 4) accessibility of target site by RNA secondary structure [14].

TargetScan [18] and miRanda [19] are popular sequence based miRNA target prediction algorithms. To run sequence based prediction algorithm over all mRNAs and miRNAs of a species requires massive computing power; Therefore, it is difficult for end users to run sequence based prediction algorithms themselves. As a result, the algorithms normally provide a web-server that incorporates the pre-computed miRNA-mRNA interaction databases for selected species, where end users can give a query (e.g. miRNA ID) to search targets of interesting miRNAs.

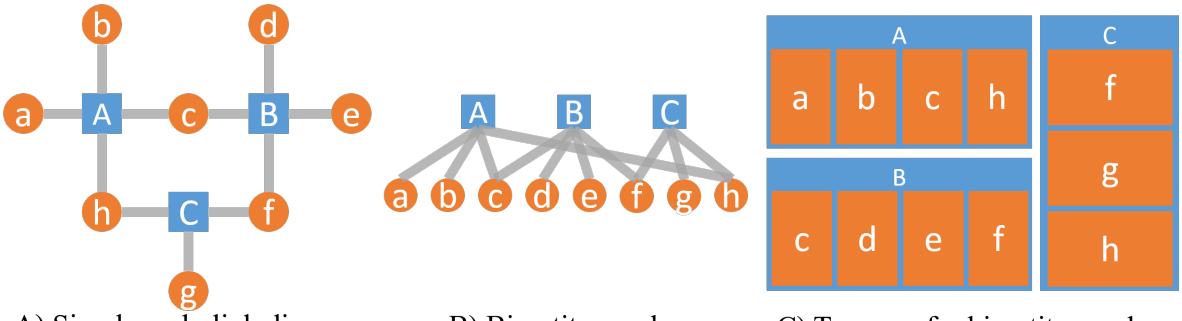
Recently, some prediction algorithms exploit whole-genome wide miRNA-mRNA expression profile data and sequence-based information such as GenMir++ [20] or Li et al. [21]. The emergence of this kind of algorithms owe to increase of miRNA-mRNA co-expression profiling experiment. As miRNAs draw more interest of researchers, many

experiments are conducted to measure mRNA and miRNA co-expression. Microarray method has been a prevalent method before NGS techniques (RNA-Seq and miRNA-Seq) methods become popular for mRNA and miRNA expression profiling method recently. As its higher accuracy, RNA-Seq are expected to replace microarrays in transcriptome expression profiling [22] [23]. According to reviews that compare miRNA-Seq with microarray [24] [25] [26], miRNA-Seq techniques outperform microarray techniques, but have higher cost.

Some analysis tools MMIA [27], miRConnX [28], or MAGIA [29] were introduced to analyze miRNA-mRNA expression profile data by miRNA prediction algorithms. They integrate various sequence-based and miRNA-mRNA expression profile based prediction algorithms for more accurate miRNA target prediction. They help users to analyze miRNA-mRNA expression data easily. They are equipped with user-friendly interface or visualize the miRNA-mRNA interactions. Especially, mirConnX [28] and MAGIA2 [30] shows miRNA-mRNA interaction network with a node-link diagram.

However, there are still a demand for more accurate miRNA target prediction with miRNA-mRNA expression data, and more effective and interactive visualizations. There is more extent to improve prediction accuracy of existing tools. The visualizations of existing tools are limited in terms of 1) scalability and 2) interaction. For example, visualizations in miRConnX [28], or MAGIA [30] have limitation for its number of possible miRNA-mRNA interactions.

Also, visualizations in existing tools support limited user interactions. It is difficult to explore the miRNA-mRNA interaction network directly in the visualization with current tools. According to Keim [31], visual data exploration follows Information Seeking Mantra [32]: Overview first, zoom and filter, and the details on demand. However, existing visualization tools for miRNA-mRNA expression profile data cannot fully support the Information Seeking Mantra [32] due to their lack of user interaction. These limitations in of existing analysis tools for miRNA-mRNA expression profile data are the motivations for developing miRTarVis.




A) Simple node-link diagram B) Bipartite graph C) Treemap for bipartite graph

Figure 1.1. Mechanism of Bipartite Treemap. A) Original node-link diagram for toy miRNA-mRNA network. B) Conversion of original network into a bipartite graph. C) Conversion of bipartite graph into a Bipartite Treemap. mRNAs that have multiple miRNA regulator (h, f) are appeared in Treemap multiple times.

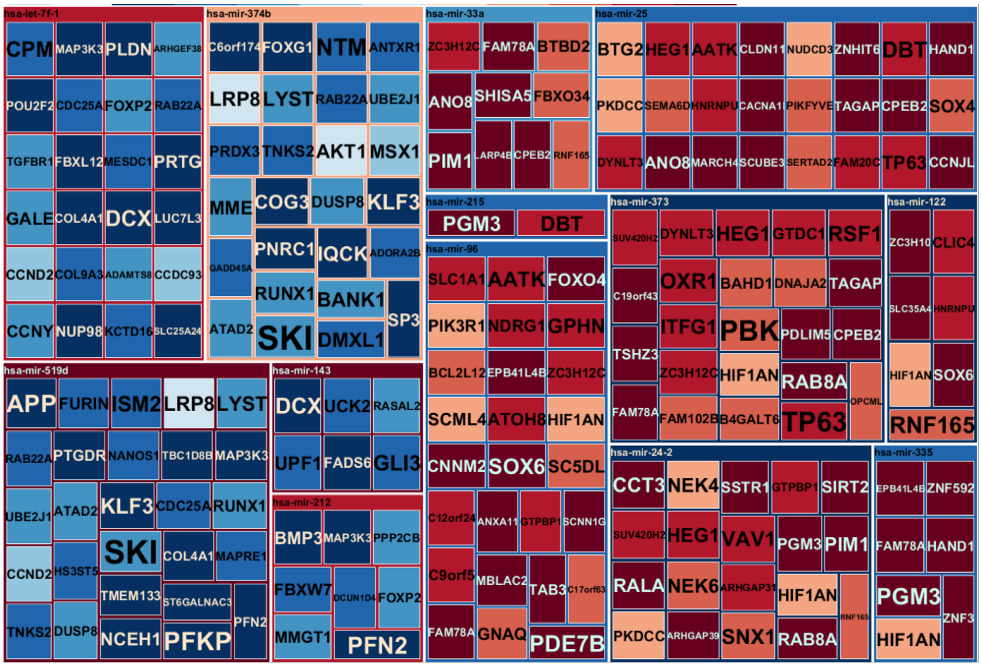
## 1.2 Main Contribution

First, we propose new visualization techniques, Bipartite Treemap and enhanced node-link diagram, for visualizing a miRNA-mRNA interaction network (Figure 1.1C, Figure 1.2A). Various tools (mirConnX [28], BioVLAB-MMIA-NGS [33], miRGator [34], MAGIA [29]) use node-link diagrams for visualizing miRNA-mRNA interaction network.

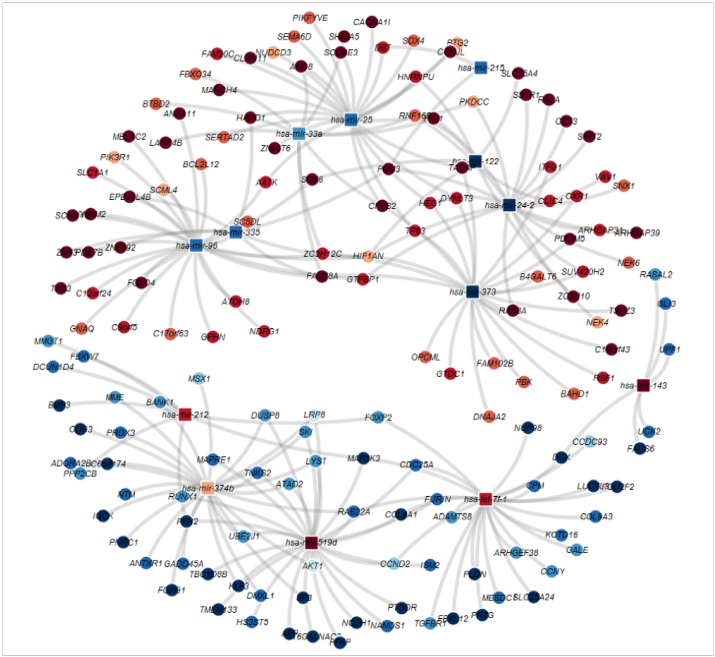
Though node-link diagram is a natural option for network visualization, it has the problem of occlusion among node and edges. The occlusion problem inhibits users to gain insight from node-link diagram visualization, especially when the target network data is overcrowded.

miRNA-mRNA interaction networks are bipartite graphs, as they have two type of nodes (miRNA and mRNA) and edges are only between different types of nodes. Furthermore, the biological observation confirms that one miRNA generally has dozens of miRNA targets, while most mRNAs have only one miRNA regulator, and a few mRNAs have multiple miRNA regulators.

Bipartite Treemap exploits this property of a miRNA-mRNA interaction network to visualize the network without occlusion problem and with high readability. Treemap [35] is an information visualization for tree data structure. We convert a miRNA-mRNA interaction network into a tree-like hierarchical data structure with two level, and convert it into Treemap (Figure 1.1). We added more user interaction to original Treemap technique to identifying mRNAs that have multiple miRNA regulators. To compensate the weakness of Bipartite Treemap that it is difficult to see structure of the network, we also use node-link diagram as existing tools do, but we enhanced node-link diagrams with some interaction techniques.




A) Treemap for bipartite graph



B) node-link diagram by KK-layout

Figure 1.2. Bipartite Treemap and node-link diagram by KK-layout for real miRNA-mRNA interaction data. Bipartite Treemap is good for reading miRNA targets. Node-link diagram is good for see overall structure of miRNA-mRNA network.

Second, we design an visual analysis tool for miRNA-mRNA expression profile data, called miRTarVis (Figure 1.3, Figure 1.4). miRTarVis is designed to help users analyze their miRNA-mRNA expression data with various miRNA target prediction algorithms on demand. miRTarVis visualizes the miRNA-mRNA interaction network by Bipartite Treemap and enhanced node-link diagram (Figure 1.2). miRTarVis supports new user interaction techniques in its visualization for exploration of miRNA-mRNA interaction network effectively.

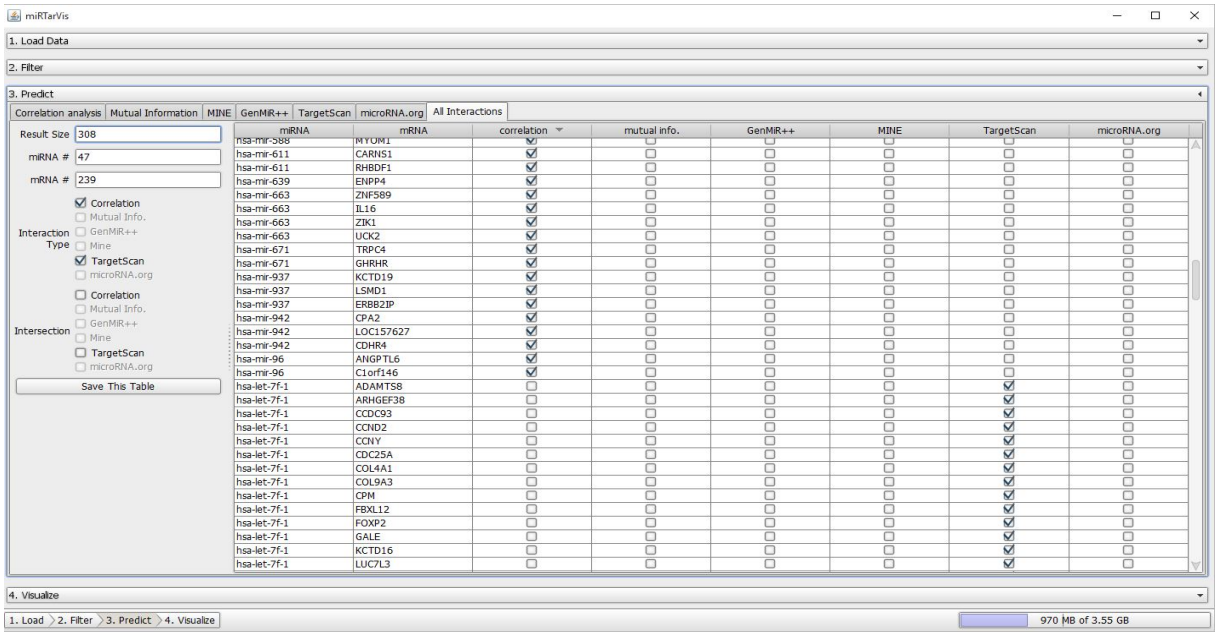



Figure 1.3. Overview of prediction menu in miRTarVis. (left) control panel for various prediction options. (right) table shows list of predicted miRNA-mRNA interaction. (up) tabs represents base prediction algorithms (e.g., Correlation analysis, TargetScan). miRTarVis combines multiple base prediction algorithms to predict miRNA targets.

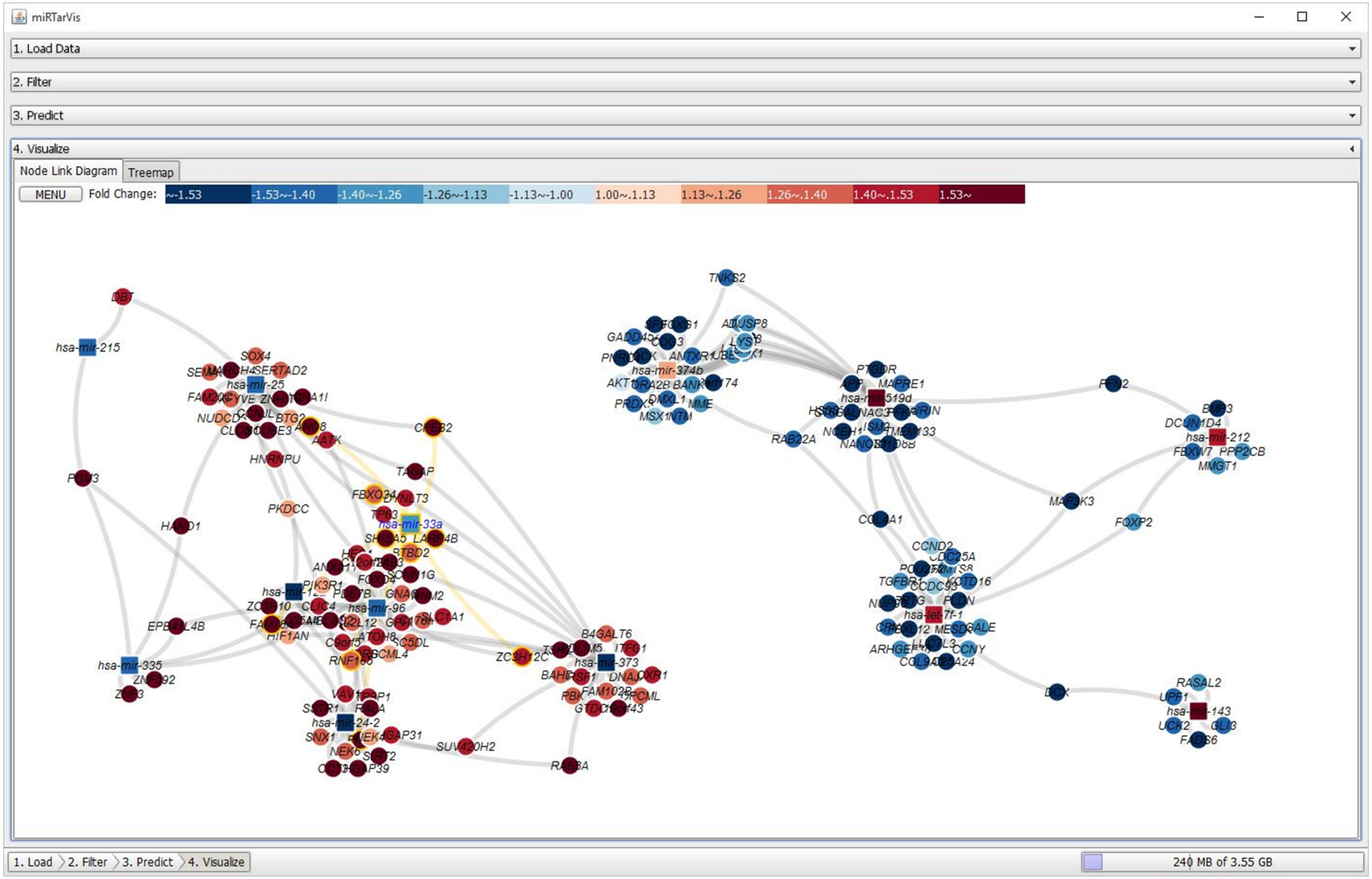



Figure 1.4. Overview of visualization miRTarVis. This figure shows the miRNA-mRNA interaction network by node-link diagram with Modified ISOM layout. Circles represent mRNAs, and rectangles represent miRNAs. Color represents fold change. Red represents up-regulation, and blue represents down-regulation. Thickness of color represents level of fold change (see the color label at the top).

In case study, our coworker applied his own miRNA-mRNA expression data from obese visceral exosomes with asthma and non-asthma. He could compare the predicted miRNA-mRNA networks from asthma patients and non-asthma patients. He used miRTarVis in independent windows, and compared the visualization side-by-sdie. He could explore his data by various analysis techniques and interactive visualizations in miRTarVis.

Finally, he could identify the important gene ACVR2B from visualizations of the predicted miRNA-mRNA interaction network in miRTarVis. In analyzing the miRNA-mRNA expression profile data, miRTarVis enables him to (1) predict putative miRNA-mRNA interaction network, (2) visualize and explore the data while changing various filtering and prediction options, and (3) confirm significant miRNAs and mRNAs that play a key role in his data.

## 1.3 Organization of the Dissertation

This dissertation is organized as follows. In Chapter 2, we will describe biological and genetic background for miRNAs. We will review two types

of techniques for microRNA predictions (sequence based and miRNA-mRNA expression profile based). After that, we will review current tools for analysis of miRNA and mRNA expression profile data and discuss their limitations and implications for designing a new analysis tool for miRNA and mRNA.

In chapter 3, we will describe our two information visualization techniques (Bipartite Treemap and enhanced node-link diagram) for miRNA-mRNA interaction network. In Chapter 4, we will describe design process, input data, analysis techniques (miRNA prediction), interface design, and implementation of miRTarVis in detail. Then, we explain two case study results in Chapter 5. We will discuss about our visualization techniques for miRNA-mRNA expression profile data and miRTarVis in Chapter 6. After that, we close this dissertation with conclusions in Chapter 7.

# Chapter 2

# miRNA target Prediction

In this section, we first review miRNA target prediction algorithms. There are two groups of prediction algorithms, sequence based and miRNA-mRNA expression profile based. We review pros and cons of each group, and discuss about necessity for meta prediction algorithms that integrate sequence based prediction algorithms with miRNA-mRNA expression profile based prediction algorithms. After that, we will comprehensively review current analysis tools that execute meta prediction from miRNA-mRNA expression profile data, and their strengths and weaknesses. Finally, we discuss design implication for a new miRNA-mRNA analysis tool from the review of current tools for developers.

## 2.1 MicroRNA Target Prediction Algorithms

MicroRNAs (miRNA) are short nucleotides (~ 22 nt) that down-regulate gene expression by degrading its target mRNAs or repressing translation of its target mRNAs [12]. The mRNAs that is down-regulated by a miRNA is called the miRNA's targets. MiRNAs have many biological functions such as early/late developmental timing or programmed cell death [36]. In addition, miRNAs play a crucial role in human disease development [37]. Therefore, there are increasing demand for analysis techniques for miRNA experiment data in Bioinformatics community.

Identifying miRNA target set is key for understanding its biological role for biologists [38]. Experiment is essential for validation of miRNA targets [16]. According to a review that summarize various experimental miRNA target identification methods [16], reporter assays, 5 RLM-RACE, or immunoprecipitation of the RISC components (e.g. HITS-CLIP) are experimental techniques that directly verifies miRNA-mRNA interactions. But it is a difficult task to identify miRNA target experimentally [39].

There are more than 35,000 miRNAs across more than 200 species, but only about 1,000 miRNA targets has been annotated as validated targets [40]. Therefore, many algorithms that can predict miRNA targets without experiment are introduced. Prediction algorithm' s goal is suggesting likely miRNA targets before a miRNA validation experiment. For example, PicTar algorithm led to the experimental validation of Mtpn gene regulation by multiple miRNAs [41]. They can restrict candidate miRNA-mRNA interactions before conducting an experiment for searching miRNA targets.

There are challenges in miRNA target prediction algorithm design. First, the size of the possible miRNA and mRNA pair is large, which means that the number of candidate solution is high. As explained in previous section, the length of miRNAs is approximately 22 bp. If we assume that length of all miRNAs is exactly 22 bp, the number of theoretically possible microRNAs is $4^{22} \approx 16 \times 10^{12}$. The number could increase because the length is not exact 22 bp.

However, fortunately, the real number of existing miRNAs are much less than millions of million. According to the latest version of miRBase

(version 21) [42], there are 2588 known unique mature human miRNAs. The number of human gene is approximately less than 30000 [43]. However, the target of miRNA is not gene, but the transcriptome (mRNA). According to a human-transcriptome database for alternative splicing [44], the number of human transcriptome is 60765. so possible. The total number of possible miRNA-mRNA pairs is approximately 157 million, which is still a big number.

Second, multiple miRNAs cooperate to regulate multiple mRNA. It is known that miRNAs control more than 30 % of human protein coding genes [45]. Multiple miRNAs regulates a mRNA [46] and a miRNA regulates multiple mRNAs [47] [48] [49]. An experiment by Wu et al. [50] confirms that a same gene is regulated by multiple miRNAs. Especially this characteristic makes it difficult to predict miRNA targets by using miRNA-mRNA expression profile data since we cannot assure that the regulation of a certain mRNA expression level is due to which miRNAs.

Third, there are many facets to consider for miRNA target prediction. It is widely-believed that sequence complementary between 5'-end region of a miRNA and conserved site in its target mRNA' s 3'-UTR is a

strong evidence of miRNA target; However, in some mRNAs, it is known that 5'-UTR plays an important role in miRNA's transcript suppression mechanism [51]. In addition, it is suggested that existence of target sites in coding domain sequence inhibits the translation [52]. According to recent publication, the mere presence of a miRNA-binding site is insufficient for predicting target regulation because targets can reciprocally control the level and function of miRNAs [53].

A perfect or near-perfect solution for miRNA target prediction problem does not exist due to the challenges described above. Visual data analysis could play an important role in this situation, where automatic bioinformatic solution could not predict all interesting miRNA-mRNA interactions from the data exactly. According to Keim et al. [54], "visualization can be used as a means to efficiently communicate and explore the information space when automatic methods fail." Visual data analysis could support data exploration and encourage knowledge discovery process by human to find hidden interesting findings in the data.

There are some review papers over miRNA target prediction algorithms [17] [39] [55] [56] [57] [58]. After studying all the reviews, we can categorize miRNA prediction algorithms into two groups: base prediction algorithm and meta prediction algorithm. Base prediction algorithm can predict miRNA target by itself, and meta prediction algorithms combine multiple base meta predictors [40]. We can regard miRNA-mRNA expression data analysis tools as meta prediction algorithm, as they combine multiple base prediction algorithms. In this perspective, mirTarVis can be categorized as meta prediction algorithm.

In this section we review base predictors, and meta predictors in the next section. We can again categorize miRNA base predictors into two groups: sequence based and miRNA-mRNA expression profile based. In the following sub-sections, we will discuss about sequence based and miRNA-mRNA expression profile based miRNA prediction algorithms in detail.

### 2.1.1 Sequence based target prediction algorithms

Sequence based target prediction algorithms uses genomic sequence information of miRNAs and mRNAs. Fan et al. [40] is one of the latest review paper about sequence based target prediction algorithms. They reviewed 38 sequence-based miRNA prediction algorithms published in last decade, and systemically evaluated 7 representative methods.

Sequence based prediction algorithms use combination of four factors: 1) sequence complementarity between miRNA's 1a) seed region or 2a) non-seed region and mRNA, 2) site accessibility for RNA-induced silencing complex (RISC), 3) evolutionary conservation of target site, 4) number of multiple target site in target mRNA. Almost all sequence (34 out of 38) based predictions uses sequence complementarity between miRNA's seed region and mRNA's 3'-UTR (Factor 1a) [40].

TargetScan [18] is the most popular sequence based miRNA prediction in terms of number of citation per year, and one of the most accurate sequence based miRNA predictions [40]. TargetScan predicts targets of miRNAs by searching conserved 8mer or 7mer sites that match the seed region of each miRNA. For a given miRNA, the web site returns

a list of predicted target genes of the miRNA in a text table. In addition, it shows how a miRNA and its target mRNA are aligned to each other.

microRNA.org [59] (www.microrna.org) provides database for miRNA targets by using the miRanda algorithm [19]. It is the third most cited sequence based prediction algorithm [40]. It uses a weighted complementary sequence score between a miRNA and a mRNA, giving a higher score for complementarity in a miRNA seed region and a lower score to complementarity in other regions. As a result, it predicts those mRNAs that match not only in the 3' end seed region of a specific miRNA but also in 3' end of a miRNA as a possible target of the miRNA. The microRNA.org web service gives a list of predicted targets sorted by the mirSVR score [60] for a given miRNA and shows alignment between a miRNA and its targets.

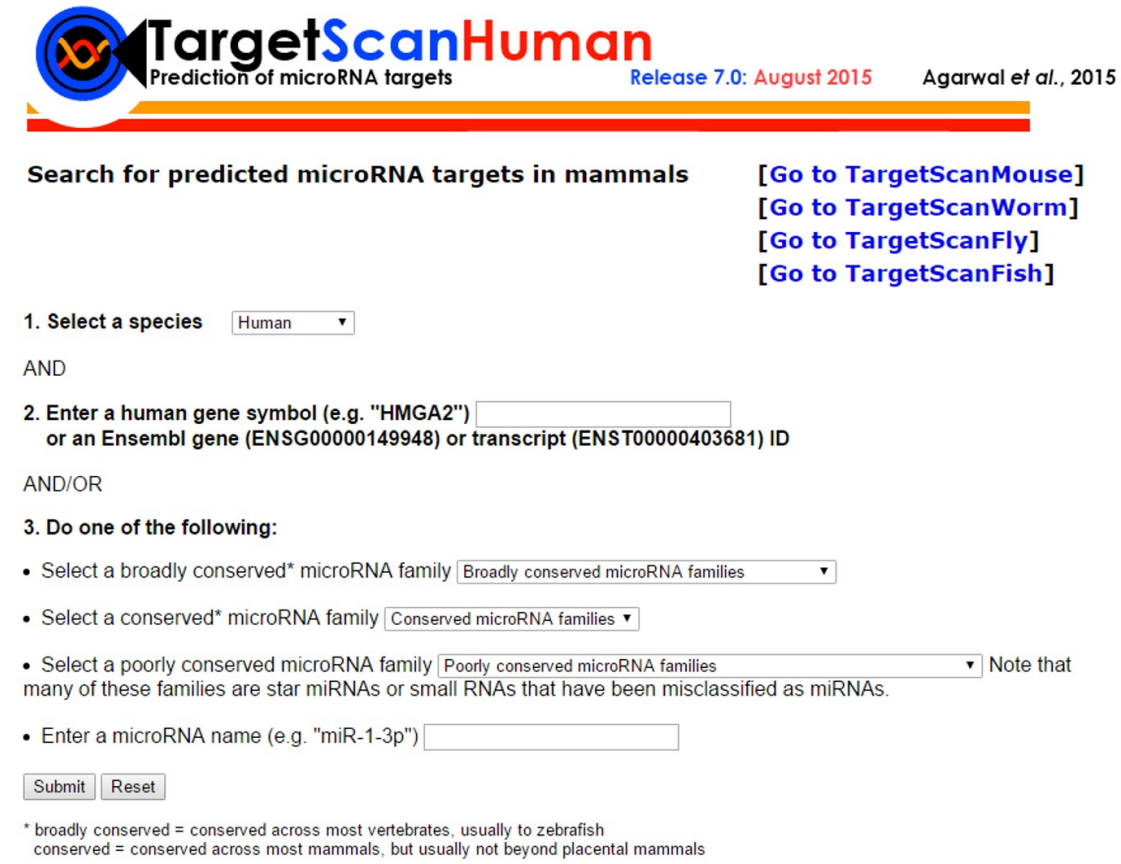
Sequence based prediction is either heuristic or empirical. Heuristic prediction uses rules that human derived from experimentally validated miRNA-mRNA interactions. The rules define the conditions about the four factors. Predicted miRNA-mRNA interaction should meet the rule. Empirical prediction uses data mining methods to predict miRNA targets.

The features for data mining is defined by the combination of the four factors, and validated miRNA-mRNA interactions are used as training data. According to Fan et al. [40], the empirical prediction yet showed lower performance than heuristic prediction.

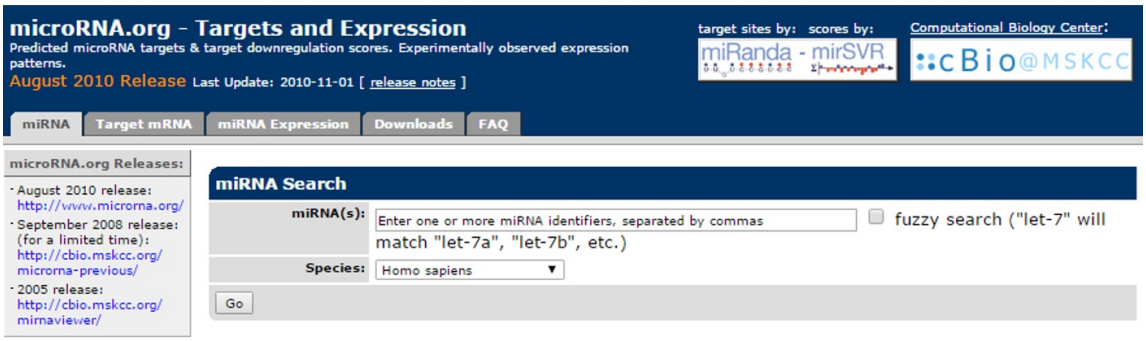
miRTarVis uses the selected sequence based algorithms (three in developmental version, two in the first version, and eight in the latest version). miRTarVis is a user of sequence based prediction algorithms. It is worth to review the prediction performance and usability of sequence prediction algorithms before using them.

Fan et al. [40] conducted a systematic evaluation over seven representative sequence based prediction. The seven representative prediction's performance is measured by comparing predicted miRNA-mRNA interactions to validated interactions from miRTarBase [61] database. The range of area under the ROC curve (AUC) and Matthews correlation (MCC) are measured for showing the overall prediction performance. AUC values across the seven prediction is between .196 and .391, as TargetScan [62] shows the highest value. MCC values are between .196 and .391, as DIANA-microT [63] shows the best.

Sensitivity ($\frac{true\ positive}{true\ positive+false\ negative}$) and specificity ($\frac{true\ negative}{true\ negative+false\ positive}$) show a propensity to have inverse relationship (if one is high, the other is low). For example TargetScan's sensitivity and specificity are .823 and .389, while PicTar's [41] sensitivity and specificity are .272 and .806. In conclusion, there was no best sequence-base prediction algorithms in universal, and each algorithm has its own strengths and weaknesses [40]. This observation shows the necessity for a meta prediction algorithm. With meta prediction, user could select and combine necessary prediction algorithms on demand.




A) TargetScan




B) microRNA.org

Figure 2.1. This figure shows interface design of web-based tools. A) TargetScan and B) microRNA.org (miRanda). Only simple queries such as “find all targets of a miRNA” or “find all miRNAs that are predicted to regulate a mRNA” is allowed in these interfaces.

Many of the sequence based predictions (24 out of 38) provides web servers where an end user can make queries for miRNA targets [40]. Sequence based prediction algorithms are usually computationally heavy because they need to run sequence alignment algorithms between set of miRNAs and set of mRNAs. Therefore, it normally is infeasible for an end user to run a sequence based prediction algorithms directly in their own local computers. For ease of use, this type of algorithms normally calculated miRNA targets for some selected species (e.g. homo sapiens) in advance.

The results are stored as a database, and a web server provides an interface for querying to the database. Figure 2.1 shows examples of such interfaces. The supporting parameters are different among interfaces. In general, users can give a miRNA ID or a gene ID with a specific species as an input parameter for the queries. The web servers return list of predicted targets of the gene or predicted regulating miRNAs, respectively.

There are some limitations in the sequence based predictions and their interfaces. The interface normally takes only one miRNA at a time.

Users cannot extract predicted miRNA-mRNA interactions among selected set of many miRNAs and mRNAs; Therefore, it is not straightforward to integrate prediction result of sequence based prediction with miRNA-mRNA expression profile data. A general analysis procedure for miRNA-mRNA expression profile data is to find some significant (e.g. differentially expressed) miRNAs and mRNAs from miRNA-mRNA expression profile data and search for miRNA-mRNA interactions among them by a prediction algorithm. However, the interfaces of sequence-based prediction algorithms are not suitable for integrating miRNA-mRNA expression profile data with prediction algorithms. This limitation suggests that analysis tool is necessary for integrating miRNA prediction algorithms and miRNA-mRNA expression data.

Many sequence-based prediction manages pre-computed database of miRNA-mRNA interactions. The databases are based on specific version of reference genome sequence (e.g. some early databases are based on hg16, and recent databases are based on hg19) and version of miRNA database (i.e. miRBase [64]). If a user' s data is based on different

reference genome version or different miRBase version with the sequence-base prediction's database, there could be some inconsistency, and the user cannot use the prediction database. The only solution is executing the prediction algorithm with the suitable version of reference genome and miRBase; However, it demands large computing power and knowledge about the execution of a prediction algorithm, which is normally not possible for a general user of sequence-base miRNA prediction.

### 2.1.2 MiRNA-mRNA expression profile based target prediction algorithms

As the cost and time for genome-wide miRNA-mRNA expression profile data decrease dramatically, more genome-wide scale miRNA-mRNA expression profile data is available. The sufficiency of miRNA-mRNA expression profile data gives motivation to improve performance of miRNA target prediction using the co-expression profile data. Predicted miRNA-mRNA interactions by sequence based prediction are inconsistent and have high false positive rate [14]. This limitation can be tackled by refining the sequence based prediction by miRNA-mRNA

expression profile based prediction algorithms. miRNA-mRNA expression profile base prediction algorithms are normally executed in three procedures: 1) starts from sequence based prediction result, 2) use a statistical or machine learning technique to score miRNA-mRNA interactions, and 3) select top-ranked miRNA-mRNA interactions. Different prediction algorithms can be characterized by the starting sequence based prediction algorithm and the statistical or machine learning algorithm they use.

We can redefine the problem of predicting miRNA-mRNA interactions as the following mathematical terms [14]. Two matrices $X_{J\times T} = [x_{jt}]$ and $Z_{K\times T} = [z_{kt}]$ denotes expression values of mRNA $j$ $(j = 1, \ldots, J)$ and miRNA $k$ $(k = 1, \ldots, K)$ in sample $t$ $(t = 1, \ldots, T)$, respectively. The putative miRNA-mRNA interactions by sequence based prediction can be denoted as a matrix $C_{J\times K} = [c_{jk}]$, where $c_{jk} = 1$ if mRNA $j$ is predicted to be a target of miRNA $k$, $c_{jk} = 0$ otherwise. The problem can be redefined as following. For given $X_{J\times T}, Z_{K\times T}$ and $C_{J\times K}$, calculate $C'_{J\times K}$ that denotes the scores for the putative interactions by sequence-based prediction.

Various strategies are used to extract significant miRNA-mRNA interactions from data. We can categorize the prediction algorithms into three groups by the type of statistical or machine learning technique they rely on. First group uses co-expression measures (e.g. relational correlation or mutual information), second group is based on linear regression (multiple or regularized), and third group exploits Bayesian methods.

Pearson correlation coefficient was used in [65] [66] [67], and Spearman correlation was used in [68] [69] to predict miRNA-mRNA interactions. Since function of miRNAs is down- regulating their target mRNAs, it is a natural choice to use only anti-correlated miRNA-mRNA interactions or miRNA-mRNA interactions that change in opposite direction (e.g., up-regulated miRNA and down-regulated mRNA) in correlation based methods. miRTarVis allows users to set up various options in prediction by correlation analysis.

Mutual information (MI) measure is an information-theoretic interpretation for computing non-linear association [70]. Maximal information coefficient (MIC) [71] is a recently introduced measure

based on mutual information. It is suggested that MIC can detect wide range of non-linear associations. MI is used in some tools for miRNA target prediction [28] [29], and MIC is the first used to predict miRNA targets using miRNA-mRNA expression in miRTarVis [72].

Multiple linear regression (MLR) searches for associations among multiple miRNAs and mRNAs simultaneously, in opposition to co-expression measure based methods, which pay attention to specific interaction. MLR is used to analyze miRNA-mRNA expression to find the effect of transfected miRNAs [73] [74]. However, when the number of samples is less than the number of miRNAs or mRNAs, the linear regression model is undermined, so the solution is not unique. In such cases, we cannot use MLR based methods.

Partial least square (PLS) methods and regularized least squares (e.g., LASSO regression or Elastic net) can be used when we cannot apply MLR due to deficiency of experiment samples [14]. Li et al. [75] use PLS method to find miRNA targets from miRNA-mRNA expression profiles of seven human colon tissues and four normal tissues. They suggest that PLS outperforms simple correlation based predictions. TaLasso

(miRNA-Target LASSO) [76] is an algorithm that uses LASSO regression for miRNA target prediction. They exert non-positive constraints to make us the property that miRNA down-regulates mRNA expression by degrading mRNA. The authors of TaLasso [76] suggest that top-ranked interactions predicted by TaLasso [76] are enriched in validated databases more than any other algorithms.

Bayesian inference is a technique that uses priori knowledge to approximate parameters in a stochastic model, and predicts by the model. GenMir++ [20] is the most popular Bayesian inference model based miRNA target predictor. Their training dataset is miRNA-mRNA expression profile data from 88 normal and cancer tissues, with 151 miRNAs and 16, 063 mRNAs. GenMir++ is evaluated by Gene Ontology enrichment. Authors of GenMir++ assume that if their prediction is accurate, miRNA targets by GenMir++ among a sequence based prediction results should have more consistent Gene Ontology annotations than randomly selected miRNA targets among the sequence based prediction results. They evaluate using TargetScanS [18] sequence based prediction algorithm, and show that result of GenMir++

has more consistent annotations than randomly selected set. In addition, they experimentally validate three miRNA-mRNA interactions among high-confident miRNA-mRNA interactions predicted by GenMir++ (miRNA let-7b regulates ACTR2, COL1A2, KPNA4). Stingo et al. [77] suggested a Bayesian graphical modeling for miRNA target prediction. Their method is different with GenMir++ in terms of regression coefficient selection and posterior inference procedure.

Two papers reviewed and compared the performance of partial set of miRNA-mRNA expression based miRNA target predictions. Muniategui et al. [14] predict miRNA-mRNA interactions by applying TaLasso [76], GenMir++ [20], Spearman correlation coefficient, and pearson correlation coefficient to five miRNA-mRNA expression datasets, and evaluate comparatively the performance of the five predictors. The five datasets [78] [79] [80] [81] [82] have different characteristic in terms of number of sample size, number of miRNAs, and number of mRNAs. Therefore, they report the comparison result with different datasets separately. They compare the performance by counting the number of validated miRNA-mRNA interactions in the top-1000 predicted miRNA-

mRNA interactions. In addition, they compare the gene-enrichment on KEGG pathway for top-1000 predicted interactions. The result shows that TaLasso [76] recovers validated miRNA-mRNA interactions best in four of the five datasets. Comparison of gene-enrichment by KEGG pathway also shows that TaLasso [76] has the best performance. However, it seems that a more systematic evaluation method is required since there is the possibility that those four datasets were over-fitted in favor of TaLasso [76]. It is necessary to apply any statistical method that can measure significance to the comparison of multiple miRNA prediction algorithms.

Le et al. [83] compare eight miRNA-mRNA expression profile based prediction algorithms. They apply Pearson correlation coefficient, IDA (value by a causal inference method) [84], MIC [71], lasso, elastic-net, Z-score [85], ProMISe [86], and GenMir++ [20] to three datasets [79] [87] [88]. They used the same evaluation criteria (enrichment of validated miRNA-mRNA interactions and KEGG pathway in predicted top-ranked miRNA-mRNA interactions) with Muniategui et al. [14]. The result shows that no technique has best performance for more than one

dataset. Though IDA measure shows the best performance on average; However, only three datasets are used, so it is hard to give much significance to the comparison result.

Result of miRNA-mRNA expression profile based miRNA target predictions is different according to input miRNA-mRNA expression profile data. This is the difference compared with sequence based prediction algorithms. The comparison conducted in the two reviews shows this characteristic. The predicted miRNA-mRNA interactions by the same technique from different datasets was different; However, miRNA-mRNA interactions predicted by sequence-based predictions are static regardless of the miRNA-mRNA expression profile data.

MiRNA-mRNA Expression data reflects the specific conditions under which the experiment is conducted. Some miRNAs or mRNAs are not functional under certain experimental conditions. Therefore, it is an expected outcome that only condition-specific miRNA-mRNA interactions are predicted by miRNA-mRNA expression profile based prediction. As a result, we could expect that miRNA-mRNA expression

based prediction algorithms are more suitable for predicting condition-specific miRNA-mRNA interactions.

The miRNA-mRNA expression profile based predictions are more likely to include indirect miRNA-mRNA interactions, since they do not take into account biologic or genetic information; Therefore, predictions could be more biologically reliable if we take into account both sequence based and miRNA-mRNA expression profile based prediction algorithms. It is not an easy task for an end user to integrate both kinds of predictions with expression profile data. To fill this gap, meta prediction algorithms emerge.

Unfortunately, there is not yet a comprehensive review that evaluate comparatively performance of existing miRNA-mRNA expression profile based miRNA target prediction algorithms. Two papers (Muniategui et al. [14] and Le et al. [83]) evaluate performance of four and eight miRNA-mRNA expression profile based prediction algorithms with five and three dataset, respectively. They use the same comparison criteria (enrichment of top-ranked predicted interactions in validated interactions and in KEGG). There is one dataset (Lu et al. [79]) that both

papers share, and the eight methods in Le et al. [83] include three methods in Muniategui et al. [14]; However, the comparison result from the same dataset is not consistent between the two papers. According to Muniategui et al., in terms of the number of validated miRNA-mRNA interactions in the top-ranked predicted miRNA-mRNA interactions, the result show that Lasso > GenMir++ > Pearson correlation coefficient. However, results from Le el al. [83] show that Pearson correlation coefficient > Lasso > GenMir++. A well-organized standard for comparative evaluation of performance miRNA-mRNA expression profile based prediction is necessary.

As there is no suggestion that guides how to combine multiple miRNA-mRNA expression profile based prediction algorithms, we try to include as many expression profile based prediction algorithms as possible in miRTarVis, and let users to select proper algorithms on demand.

## 2.2 Analysis Tools for Integrated Analysis of miRNA and mRNA

As discussed in previous section, sequence based and miRNA-mRNA expression profile based prediction algorithms have pros and cons. A couple of analysis tools are introduced to overcome the limitations of the prediction algorithm. They normally 1) use meta prediction algorithm, 2) integrate differential analysis for genes and miRNAs with miRNA target prediction, 3) help users use miRNA target prediction more easily, or 4) use visualization for miRNA-mRNA interaction network. In this section, we will review these analysis tools and discuss their strengths and weaknesses, and discuss about suggestions for design of a new analysis tools of miRNA-mRNA expression profile data. As long as our knowledge, there is not yet a comprehensive review for miRNA-mRNA analysis tools.

We could find seven tools for integrated analysis of miRNA-mRNA. We summarize them in terms of nine properties (publication year, input data type, validated miRNA target database used, sequence based

prediction used, miRNA-mRNA expression profile based prediction used, integration method for combining multiple miRNA-mRNA interactions, support of differential analysis, visualization and evaluation). Table 2.1 shows the result.

MMIA [27], miRConnX [28], Magia [29], ProteoMirExpress [89], and BioVLAB-MMIA-NGS [33] integrated miRNA-mRNA expression profile data with sequence-based prediction algorithms. Especially, ProteoMirExpress [89] uses protein expression profile as well as miRNA and mRNA expression profiles. Only BioVLAB-MMIA-NGS [33] can take raw sequencing data as input. Alignment of raw sequencing data (reads) to reference genome is a computationally expensive task. BioVLAB-MMIA-NGS [33] has a goal to be a comprehensive system for NGS data. It is a big system that uses a cloud computing server with large computing power. miRGator [34] and ComiRNet [90] do not take miRNA-mRNA expression profile data as input. miRGator takes a miRNA as input, and ComiRNet takes list of miRNAs and mRNAs as input.

Table 2.1.Summary of seven analysis tools for integrated analysis of miRNA and mRNA. 'Year' represents the year of publication. 'Input data' represent type of input data for the tool. 'Validated' represents validated miRNA target databases the tool uses. 'Sequence based' represents sequence based miRNA prediction algorithms the tool uses. 'Expression based' represents which miRNA-mRNA expression profile based algorithms the tool uses. 'Integration method for multiple predictions' represents the method that the tool uses for integrating multiple miRNA-mRNA interactions. 'Differential analysis' represents differential analysis techniques the tool uses. 'Visualization' represents how and what the tool visualizes from the analysis results. 'Evaluation' represents the evaluation method of the publication for the tool.

| Name | Ref. | Year | Input Data | Validated | Sequence based | Expression based | Integration method for multiple predictions | Differential analysis | Visualization | Evaluation |
|---|---|---|---|---|---|---|---|---|---|---|
| MMIA | [27] | 2009 | * custom miRNA-mRNA expression profile<br>* list of miRNAs and mRNAs | No | TargetScan<br>PicTar<br>PITA | No | intersection | t-test | Heatmap of miRNAs expression profile. | case study to data from previous work |
| MAGIA | [29] | 2010 | * custom miRNA-mRNA expression profile data<br>* list of miRNAs and mRNAs | No | TargetScan<br>PicTar<br>PITA<br>miRanda<br>RNAHybrid | Spearman Cor.<br>Pearson Cor.<br>MI<br>GenMiR++ | * union/intersection to sequence based predictions<br>* intersection of sequence based and expression based predictions | meta-analysis (based on Empirical Bayes test) | node-link diagram of miRNA-mRNA interaction network by GraphViz | case study to data from previous work |
| miRConnX | [28] | 2011 | custom miRNA-mRNA expression profile data | TarBase<br>miRecord | TargetScan<br>PicTar<br>PITA<br>miRanda<br>RNAhybrid | Spearman Cor.<br>Pearson Cor.<br>Kendall Cor.<br>MI | union | No | node-link diagram of miRNA-mRNA interaction network by Cytoscape | case study to data from previous work |
| miRGator 3.0 | [34] | 2012 | a miRNA | miRecords<br>TarBase<br>mirTarbase | TargetScan<br>PicTar<br>PITA<br>miRanda<br>microcosm<br>miRDB | No | union | No | * node-link diagram for one miRNA and its target<br>* scatter plot of miRNA-mRNA expression in pre-selected public dataset | No |
| ProteoMirExpress | [89] | 2013 | miRNA-mRNA-proteomic expression profile data | miRecords<br>TarBase<br>starBase | TargetScan<br>PicTar<br>PITA<br>miRanda | Pearson Cor. | * union validated and sequence based predictions<br>* intersection between validated + sequence based prediction and expression based prediction | No | node-link diagram of miRNA-mRNA-protein interaction network by Cytoscape | case study to data from previous work |
| BioVLAB-MMIA-NGS | [33] | 2014 | * NGS sequencing data (RNA-Seq, miRNA-Seq)<br>* miRNA-mRNA expression profile data by microarray | No | TargetScan<br>PITA<br>miRanda<br>PMTED | No | union<br>intersection | DESeq<br>Cufflinks<br>Limma | node-link diagram of miRNA-mRNA interaction network by Cytoscape | No |
| ComiRNet | [90] | 2015 | list of miRNAs and mRNAs | No | TargetScan<br>PicTar<br>PITA<br>miRanda<br>microCosm<br>Diana-microT | No | ensemble learning approach | No | No | No |

TargetScan [18] and PITA [91] are the most widely-used sequence based prediction algorithm. All seven tools use them. There are two strategies how tools use miRNA-mRNA expression profiles data in miRNA prediction. The first strategy is to use the data for miRNA-mRNA expression profile based prediction algorithm. The second strategy is to use the data for differentially expressed (DE) analysis to find DE genes (DEG) and DE miRNAs (DEmir).

miRConnX [28], Magia [29], ProteoMirExpress [89] use the first strategy. Pearson correlation coefficient is used by all the three tools. Spearman correlation and MI are used by two tools. Interestingly, none of the tools uses multiple linear regression or regularized regression techniques yet.

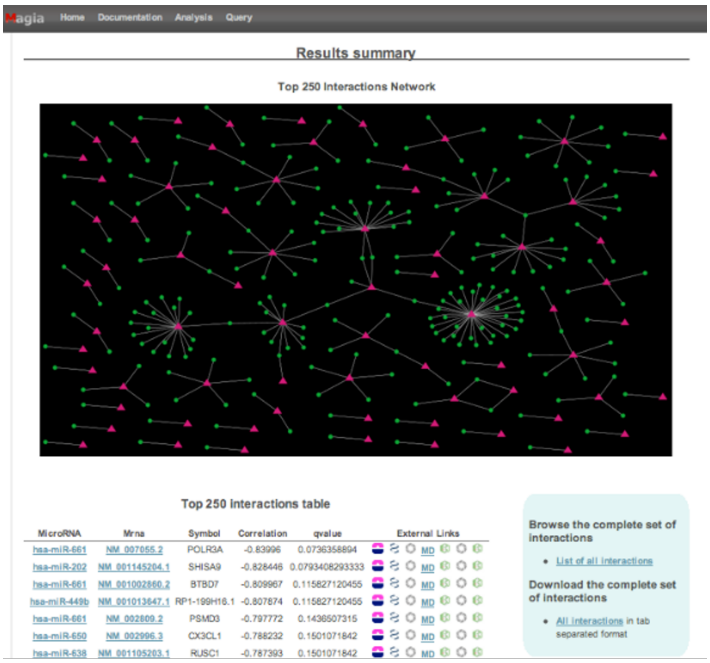
MMIA [27], Magia [29], BioVLAB-MMIA-NGS [33] use the second strategy. They find DEGs or DEmirs first, and search miRNA-mRNA interactions among the DEGs and DEmirs. MMIA [27] uses simple t-test. Magia [29] uses its own "meta-analysis" based on empirical Bayes test [92]. BioVLAB-MMIA-NGS [33] uses advanced DE analysis for

sequencing data (DESeq [93] and Cufflinks [94]) and microarray data (Limma [95]).

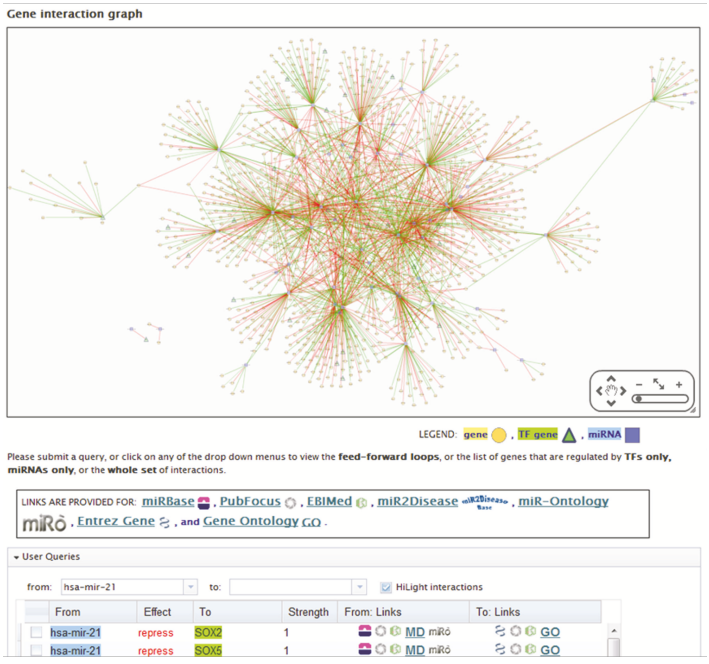
Except ComiRNet [90], all analysis tools support visualization that shows its analysis result. MMIA [27] shows a heat map of DEmirs. All the other tools show predicted miRNA-mRNA interaction network in node-link diagram by Cytoscape [96] or GraphViz (http://www.graphviz.org/) (Figure 2.2); However, the visualization of these tools have some limitations.

The visualization module of each tool is separated with the analysis module of the tool. All tools use external visualization agent (Cytoscape or GraphViz) to visualize their miRNA regulation network. This system structure is suitable for visualizing the result, but it is not suitable for making query or filtering uninteresting elements directly on the visualization. Direct manipulation to visualization is essential for visual analysis of data; However, existing tools do not support enough user interaction for visual analysis. miRTarVis focuses this point, and support simple basic user interaction in its visualization.

Four of the reviewed tools evaluated their efficacy in their papers. All the four papers apply their tools to previously published experiment data, and show that their tools re-confirm the validated results. For evaluation of miRTarVis, we also conduct a case study by applying mirTarVis to experiment data and confirming whether it is good at giving insight to its user.




A) Visualization in MAGIA




B) Visualization in mirConnX

Figure 2.2. miRNA-mRNA network visualization in A) MAGIA and B) miRConnX. MAGIA has a limitation on the maximum number of interactions at 250. Both MAGIA and miRConnX have the problem that they needs to reconstruct the visualization when a user changes a prediction options. miRConnX supports panning and zoom interaction.

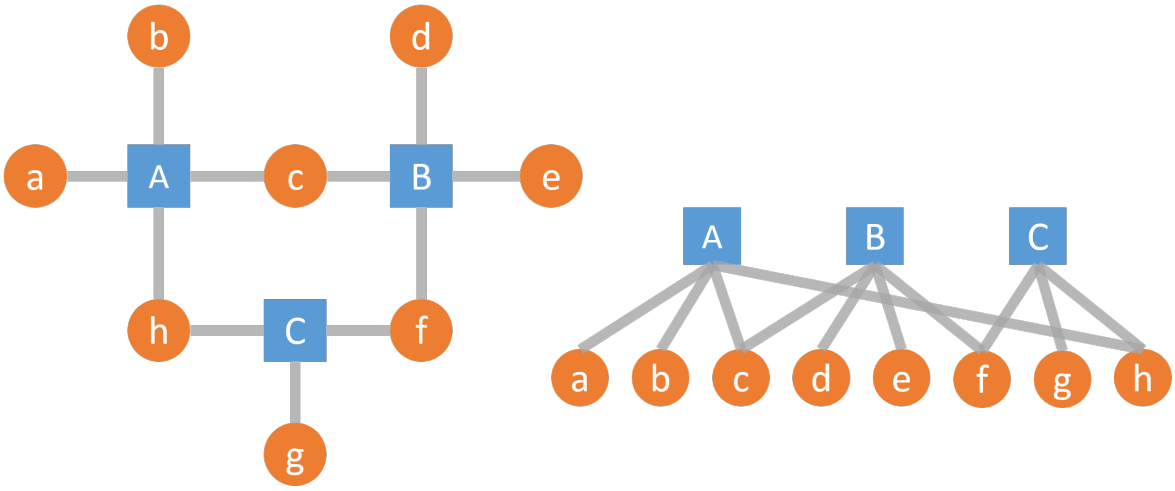
# Chapter 3

# Bipartite Treemap and Enhanced Node-Link Diagram for miRNA-mRNA Interaction Network

We propose a new visualization technique for miRNA-mRNA interaction networks, "Bipartite Treemap." Bipartite treemap is to present miRNA-mRNA interaction networks without node or edge occlusion. In addition Bipartite treemap is to present significance of the miRNA-mRNA interactions (e.g. p-value or number of supporting prediction algorithms) more effectively. Node-link diagrams (Figure 3.1A shows a node link diagram for example toy data, Figure 3.4A and Figure 3.4B show a node link diagram for real mRNA-mRNA expression data) can show overall structure of a miRNA-mRNA interaction network, but they have occlusion problem for large data. Bipartite Treemap (Figure 3.1C

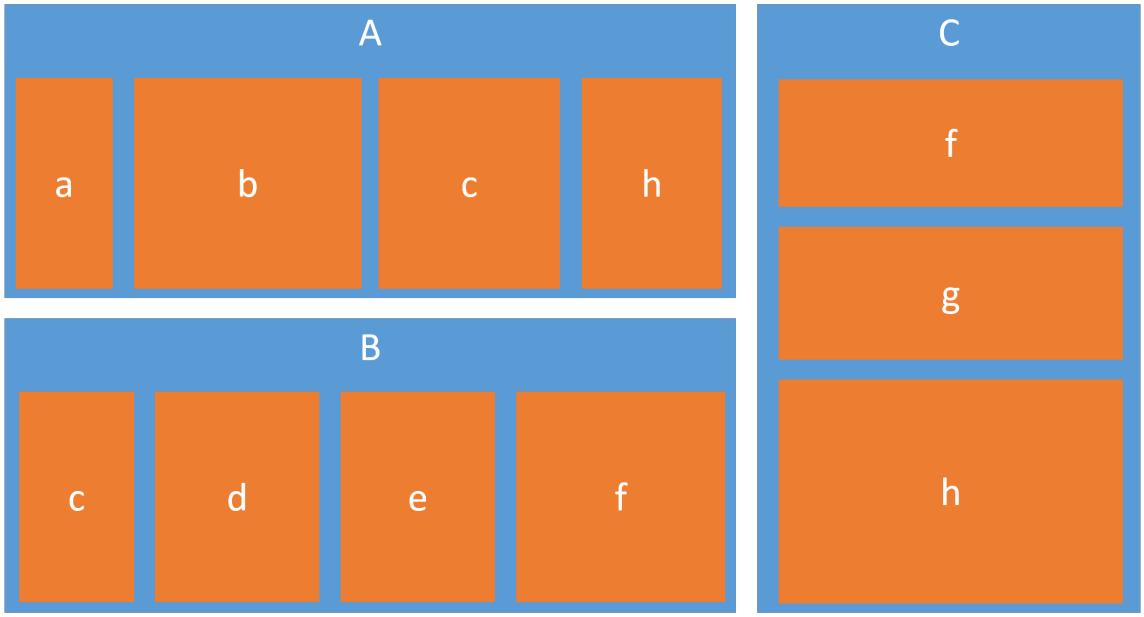
for toy data, Figure 3.4C for real data) expands Treemap [35] [97] visualization for reducing occlusion problem, encoding significance of interaction and enhancing label readability.

In addition, we propose user interaction techniques and user-adjustable graph layout algorithms for node-link diagram of miRNA-mRNA interaction to enhance users' data exploration in miRNA-mRNA interaction network. There are miRNA-mRNA analysis tools that provides node-link diagram for miRNA-mRNA interaction network, but they support limited user interaction and use fixed graph layout algorithm. Users can make queries or see the details directly in our node-link diagrams. Our node-link diagram also shows the network with various graph layout algorithms on users' demand.




A) Simple node-link diagram

B) Bipartite graph




C) Bipartite treemap

Figure 3.1. Concept of Bipartite Treemap. miRNA-mRNA interaction network (a) is a bipartite graph. Bipartite graphs can be visualized like (b), where two group of nodes are aligned in parallel; However, large occlusion among edges occurs. Bipartite Treemap (C) represent bipartite graph of miRNA-mRNA interaction network as Treemap. The significance measure of miRNA-mRNA interactions, which is difficult to represent in node-link diagrams, is represented by the size of rectangle size and easily identified by users.

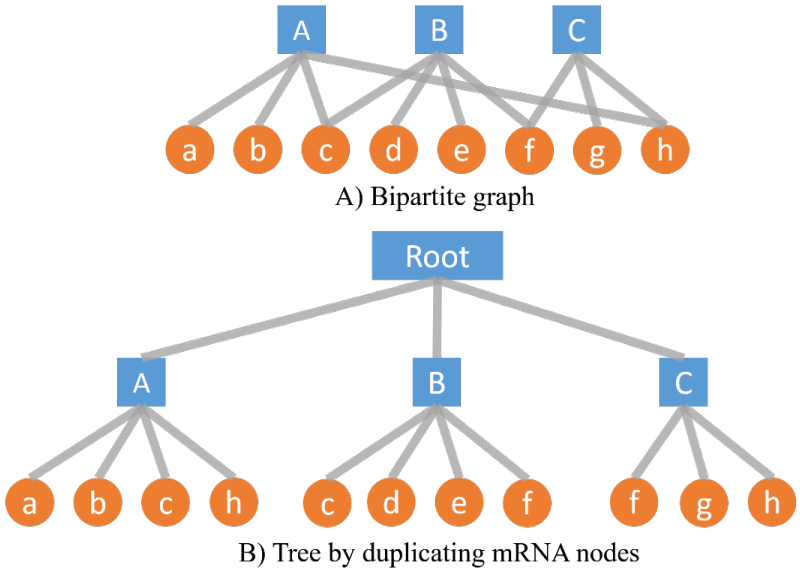



Figure 3.2. Conversion of miRNA-mRNA interaction bipartite graph into tree. Uppercase letters represent miRNAs and lowercase letters represent mRNAs. By duplicating mRNA nodes with multiple miRNA regulators, a bipartite graph (A) is converted into a tree with two levels (B).

## 3.1 Visual representation of Bipartite Treemap

Treemap [35] is originally a visualization technique for showing hierarchical tree data structure. Therefore, to visualize a network by Treemap, we need to convert a graph to a tree by removing cycles in the graph. MiRNA-mRNA interaction network is bipartite. There are two types of nodes (miRNA vs. mRNA), and there is no link among nodes of common type. We can convert a bipartite graph of miRNA-mRNA

interactions to a treemap by duplicating multiple targeted mRNA nodes (Figure 3.2). Then, we apply squarified treemap algorithm [97] to the tree. We call the generated treemap as Bipartite Treemap.

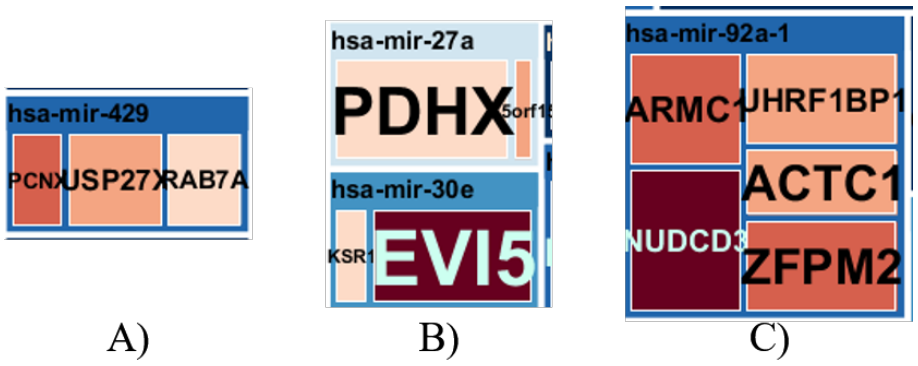
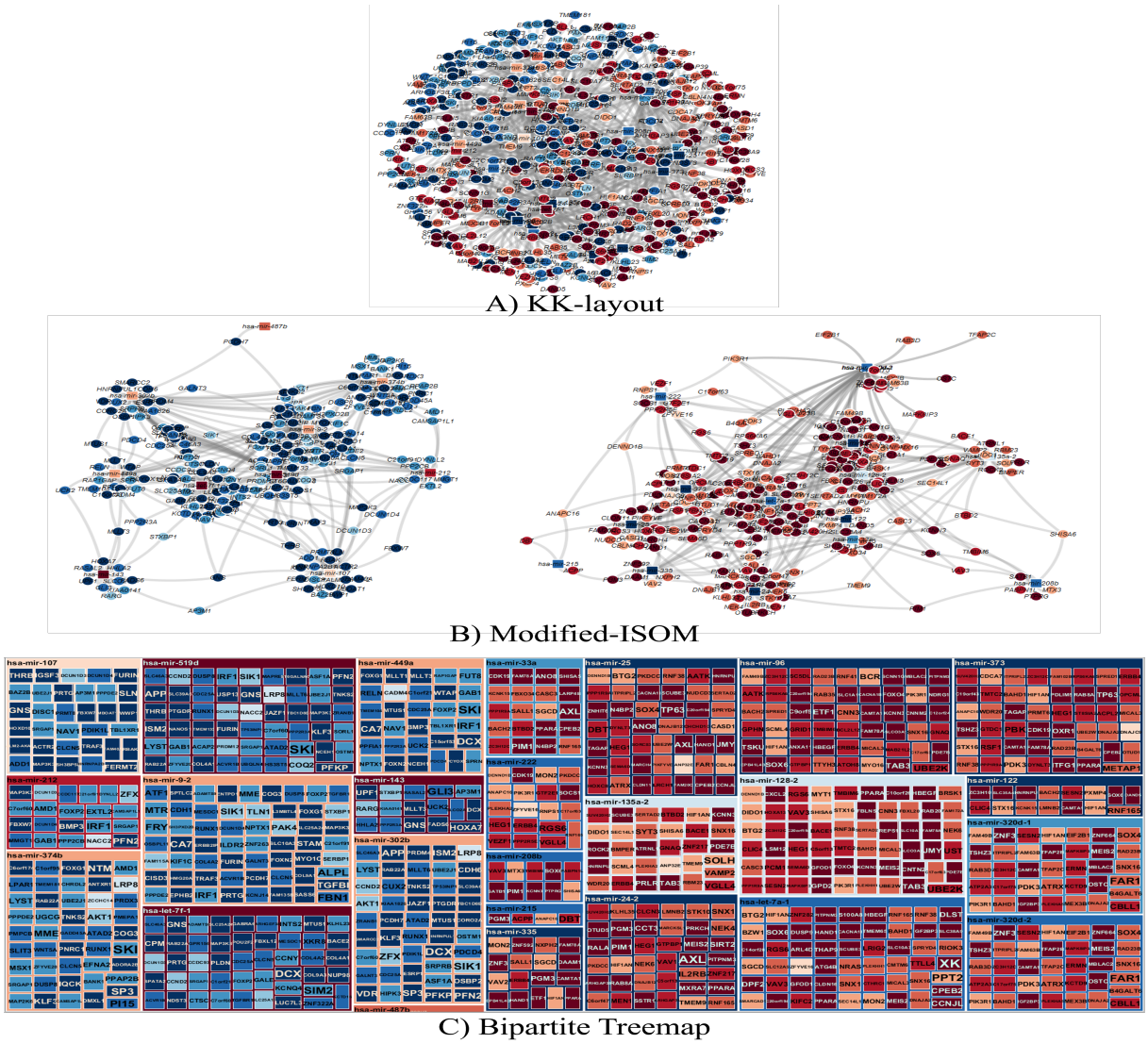



Figure 3.3.These figures shows enlarged parts of Bipartite Treemap that visualize a real miRNA-mRNA interaction network. The size of mRNA nodes encodes significance of prediction that is calculated in prediction analysis. For example, in B), the interaction between hsa-mir-27a and PDHX is more biologically significant than the interaction between hsa-mir-27a and 5orf1.

The color of node in Bipartite Treemap encodes represent the fold change value. The hue (red or blue) encodes direction of fold change, and the saturation (thickness of the color) encodes the level of fold change. When the fold change is not available (samples are not in two groups), the color of miRNA is orange, and color for mRNA is dark blue.

In node-link diagram, it is difficult to encode significance of miRNA-mRNA interactions. It could be represented by the edges' color or

thickness, but it is not easy for a user to identify visual properties of edges. In Bipartite Treemap, the size of mRNA nodes represents the biological significance of the miRNA-mRNA interaction calculated during miRNA prediction analysis (Figure 3.3).



A) KK-layout

B) Modified-ISOM

C) Bipartite Treemap

Figure 3.4. Bipartite Treemap and two node-link diagrams with KK-layout and modified-ISOM graph layout algorithms for the same given area.

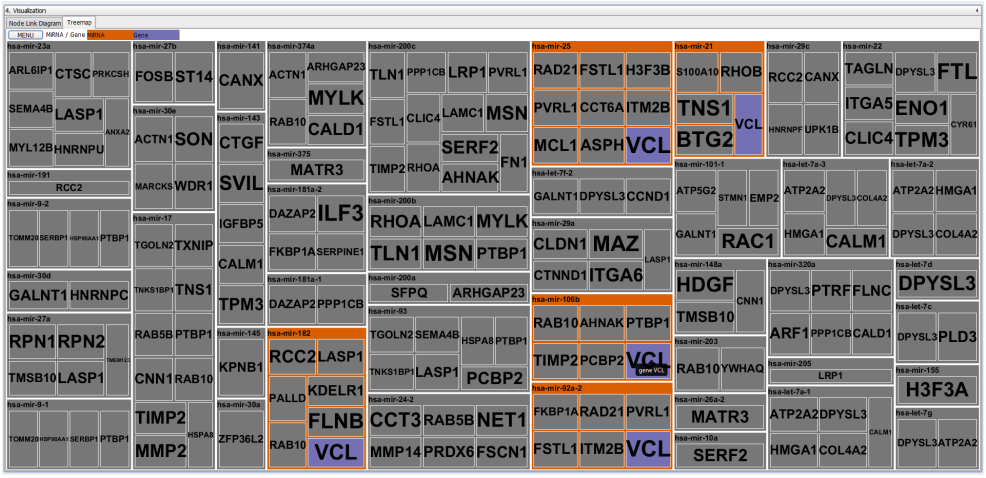



Figure 3.5. When users move mouse cursor into an mRNA of interest, other nodes that represent the identical mRNA and the miRNAs predicted to regulate the mRNA are highlighted in the Treemap visualization. All the other mRNA and miRNAs are faded out by grey color.

In this way, it is easy to see targets of miRNA in Bipartite Treemap than node link diagram. Treemap is a space-filling technique, so nodes in Bipartite Treemap take larger area than any graph-layout algorithm. This property helps users to read labels for nodes. Figure 3.4 shows two node-link diagrams by Kamada-Kawai (KK) and modified ISOM layout algorithms and Bipartite Treemap for the same given area. Bipartite Treemap fully uses the given space, but KK layout only part of the space because it makes all nodes in a virtual circle. Modified ISOM layout is

better than KK layout, but it is still more difficult than Bipartite Treemap to read node labels in node-link diagrams.

However, the limitation of Bipartite Treemap is that the same mRNA nodes can come out multiple times. Predicted miRNA-mRNA interaction network has the property that most mRNAs have one miRNA regulator; However, there are still some mRNAs that have multiple miRNA regulator, and it is hard to see identify mRNAs in Bipartite Graph. We introduce a simple interaction technique to show mRNAs with multiple associated miRNAs in Bipartite Treemap. When a user hovers mouse cursor over a mRNA, duplications of the mRNA is pointed out (Figure 3.5).

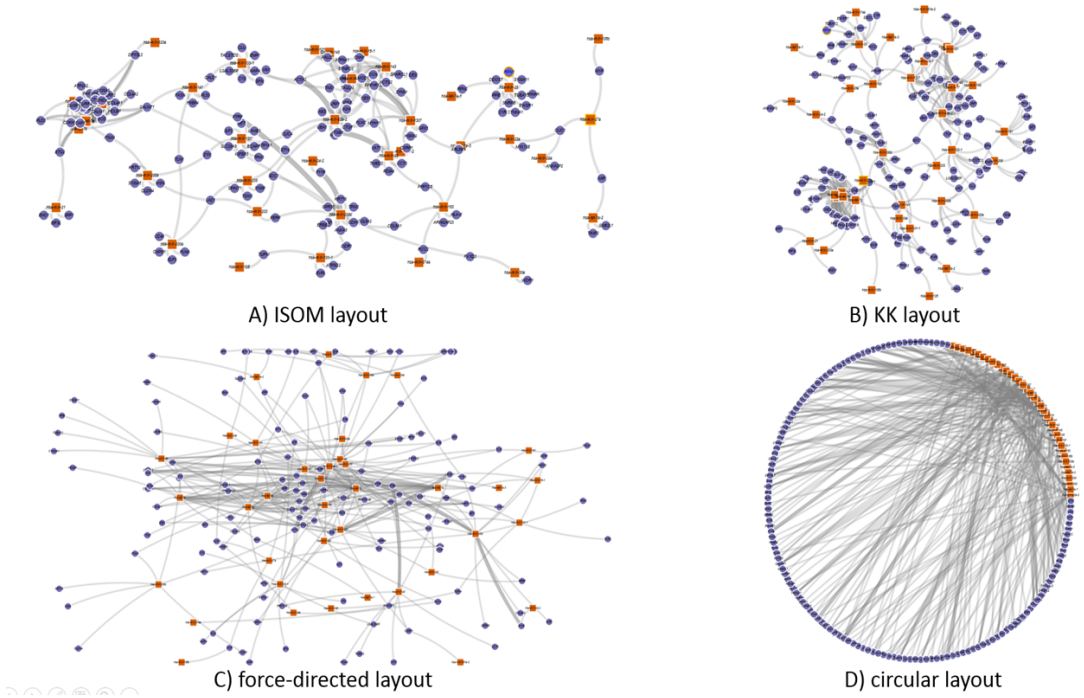



Figure 3.6. Four layouts for node-line diagram of miRNA-mRNA interaction network.

## 3.2 Node-link Diagram with Enhanced Interaction and Various Graph Layouts

Bipartite Treemap has no occlusion and better label readability, but node-link diagram has its strength in showing overview structure of miRNA-mRNA network better and discovering important bridging node in the network; Therefore, we also use node-link diagram as well as Bipartite Treemap to visualize miRNA-mRNA interaction network. Three previous tools [33] [29] [28] show miRNA-mRNA interaction

network in node-link diagram. They have the limitation that 1) they use only force-directed layout algorithms, and 2) they only support simple zoom and pan interaction.

We introduce enhanced node-link diagram to overcome the limitation. First, we introduce various layout algorithms suitable for miRNA-mRNA interaction network. After applying available algorithms, we heuristically identify that ISOM layout [98], and KK layout [99] algorithms are aesthetically satisfying, so we use these layout algorithm. Furthermore, since its popularity, there are still demands for force-directed layout and circular layout algorithm. We enable users to flexibly select one of the four layout algorithms according to their demands (Figure 3.6).

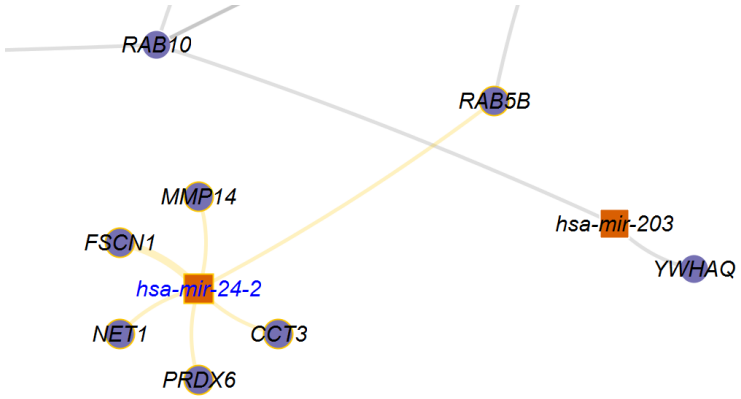



Figure 3.7. This figure shows a zoomed-in region of the node link diagram which miRTarVis' modified ISOM layout is applied to. Among predicted targets of hsa-mir-24-2, mRNAs connected to hsa-mir-24-2 with a single link are placed around it. In original ISOM layout, they were placed at the same position, so it is difficult to recognize them because of occlusion.

ISOM layout algorithm is useful to discover the structure of a predicted miRNA-mRNA interaction network because it only considers connection between nodes. A mRNA that is regulated by multiple miRNAs comes in the middle of the miRNAs in ISOM layout; However, original ISOM layout places mRNAs that is regulated by one common miRNA at the same position, so serious occlusion among multiple mRNAs usually occurs. To cope with this problem, we modify ISOM layout so that mRNAs regulated by the only same miRNAs are placed around the miRNA evenly (Figure 3.7).

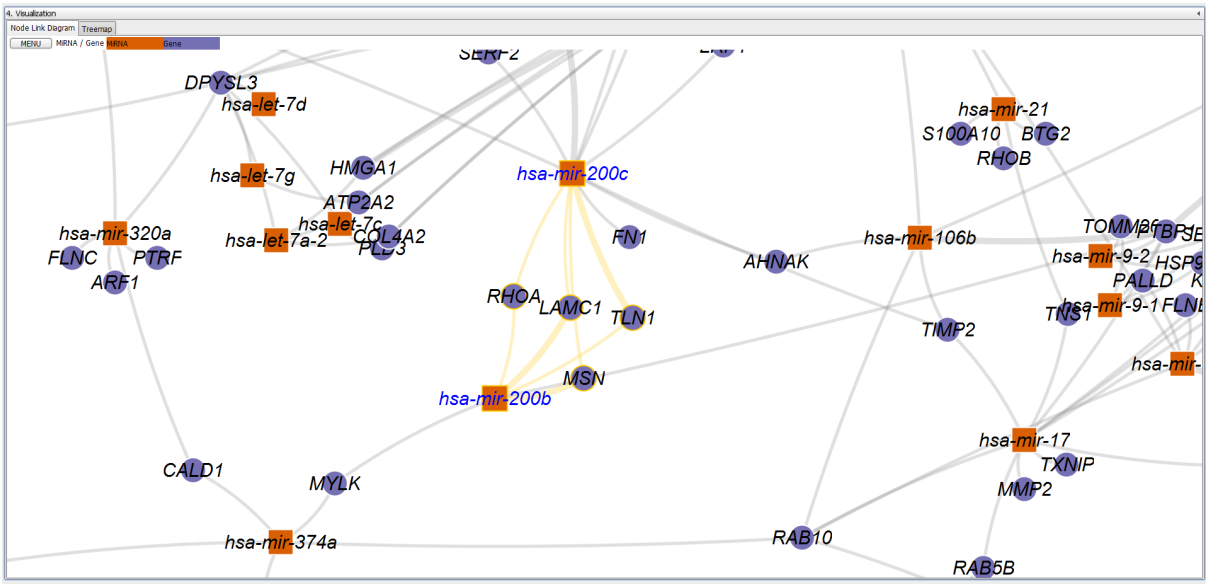



Figure 3.8. When a user selects multiple miRNAs, the co-targets of the selected miRNAs are highlighted. In this figure, two miRNAs (hsa-mir-200c and hsa-mir-200b) are selected, and only edges that link the selected miRNAs and their 4 co-targets (PHOA, LAMC1, TLN1, MSN) are highlighted in orange color.

We support simple zoom and pan interaction. The node-link diagram supports intuitive zooming in/out interaction by mouse wheel, and panning interaction by dragging with the right mouse button. To show detail on users' demand, double-clicking on a link shows a scatterplot between the corresponding miRNA and mRNA pair. For easy exploration in the visualization, when users click on a node in a node-link diagram, connected links and nodes are highlighted by bright yellow color. If a

user selects multiple miRNAs (clicking with shift key pressed), miRTarVis highlights their common targets (Figure 3.8). This feature is valuable because commonly-targeted mRNA could play an important role in a miRNA-target network.

Through a context menu, miRTarVis provides external links to miRBase [42] and miR2Disease [37] for miRNAs and to NCBI and GeneCards [100] for mRNAs. The ISOM and KK layouts of miRTarVis present an effective overview of miRNA regulatory network, but they have a common limitation of node occlusion with excessive nodes and links. Users can alleviate this problem by relocating a node manually (by dragging and moving nodes) to mend for a clean node-link diagram exported in their publication.

## 3.3 Interfaces and Interaction Design for Bipartite Treemap and Enhanced Node-Link Diagram

The goal of Bipartite Treemap and enhanced node-link diagram is visualize miRNA-mRNA expression profile data. They are not a general purpose visualizations for network data. They are specifically confined

to visualize predicted miRNA-mRNA regulatory network data. Therefore, The visualizations are closely related to analysis of miRNA-mRNA expression data. A predicted miRNA-mRNA regulatory network is the result of an analysis of miRNA-mRNA expression data. Therefore, we will explain how our new visualization techniques are integrated with the interfaces for analysis of miRNA-mRNA expression profile data.

In addition, Bipartite treemap and enhanced node-link diagram support various interaction techniques for helping researchers to analyze their miRNA-mRNA expression profile data more articulately. The real usefulness of these new visualization techniques can be exposed only when the visualizations are used with their rich user interaction techniques. Therefore, in this section, we also will discuss how our new visualization techniques, Bipartite Treemap and enhanced node-link diagram, are designed to use for analyzing miRNA-mRNA expression profile data with our interface.

When we first discussed about the design of useful visualization techniques and interface design for bioinformatic analysis of miRNA-mRNA expression profile data with our collaborating miRNA-mRNA

researchers, at the early stage of our visualization design process, we designed our interfaces and visualization techniques to perform all necessary procedures for analysis in one large screen space (Figure 3.9). Our collaborators said that it is important to load, process, and visualize the data in one screen space for a visualization technique for miRNA-mRNA expression profile data. The first prototype interface design for visualization was the result of reflection of our collaborators' comments.

In our first design prototype, all interfaces for analysis procedures including loading data, filtering data, predicting miRNA targets, and visualizing predicted miRNA-mRNA prediction network are appeared simultaneously. Interface for loading miRNA expression profile data is on left, interface for loading mRNA expression profile data is on right, interface for predicting miRNA targets is on top, and interface for visualization is on center. This design prototype is reflecting natural data processing procedures for miRNA-mRNA expression profile data. miRNA expression data is located on left, while mRNA data is located on right. The main center panels that are for integrated analysis of data from left and right panels, that is, miRNA and mRNA data. The top region

of main center panel is for prediction of miRNA-mRNA expression profile data. The bottom panel is the visualization panel, which shows the result of execution of analysis that uses the parameters specified in the top panel.

Users can adjust the various prediction parameters according to their own prediction goal and expected performance. If they want the miRNA target prediction to find lightly predicted miRNA-mRNA interactions, they can make the parameters grow lower values. If they want to find tightly predicted miRNA-mRNA interactions, they can make the parameters grow higher by moving the knot of the sliders into the right. Also, they can combine multiple predictions by checking multiple predictions in the panel.

However, that design lacked guidance or affordance about how to perform the analysis procedure in order and required a large screen space. If the novice user sees the location of the panels of the first prototype design, they can be confusing because they do not get the affordance about where to start the analysis by loading the miRNA and mRNA expression profile data. Therefore, we had to confine the visualization

design to make users follow the analysis procedures for miRNA-mRNA expression profile data. The best option for confining users to follow pre-defined analysis procedure is to give them limited degree of freedom by providing limited mode change options in the interface.

Furthermore, the screen space for visualization becomes small in this visualization. For example, for analyzing human miRNA-mRNA expression data, there are approximately 2000 miRNAs and 20000 mRNAs for human. To show users the detailed information about the input miRNA and mRNA expression data, the left and right panels, which are for miRNA and mRNA expression profile data input repectively, takes relatively large spaces. This results in the reduction of space of visualization for miRNA-mRNA network, which is the most important and essential part of the interface.

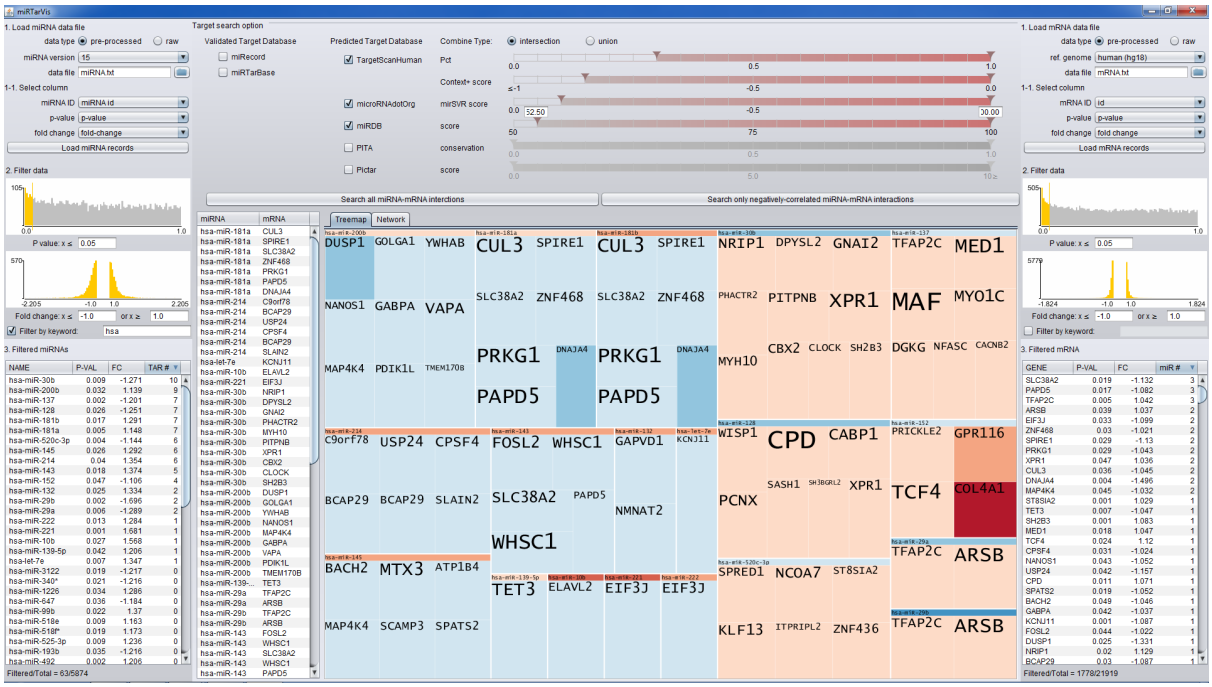


Figure 3.9. Early interface design for visual analysis of miRNA-mRNA regulatory networks. Left panels are for miRNA, right panels are for mRNAs, and the center panels are for integrated analysis of miRNA and mRNA expression profile data.

To address these problems, we redesigned the interfaces for the Bipartite treemap and enhanced node-link diagram in accordance with load-filter-predict-visualize procedure for miRNA-mRNA expression profile data (Figure 3.10). We adopted a simple systematic foldable accordion metaphor, where only one selected menu item is open while all others are collapsed. This is more intuitive to users because the order of the menu items well-matches the order of the analysis procedure.

Through this foldable accordion interface, users can focus on the current step of the procedure while using the screen space more effectively even in a small-size screen (minimum available for a screen resolution of 1280 x 720).

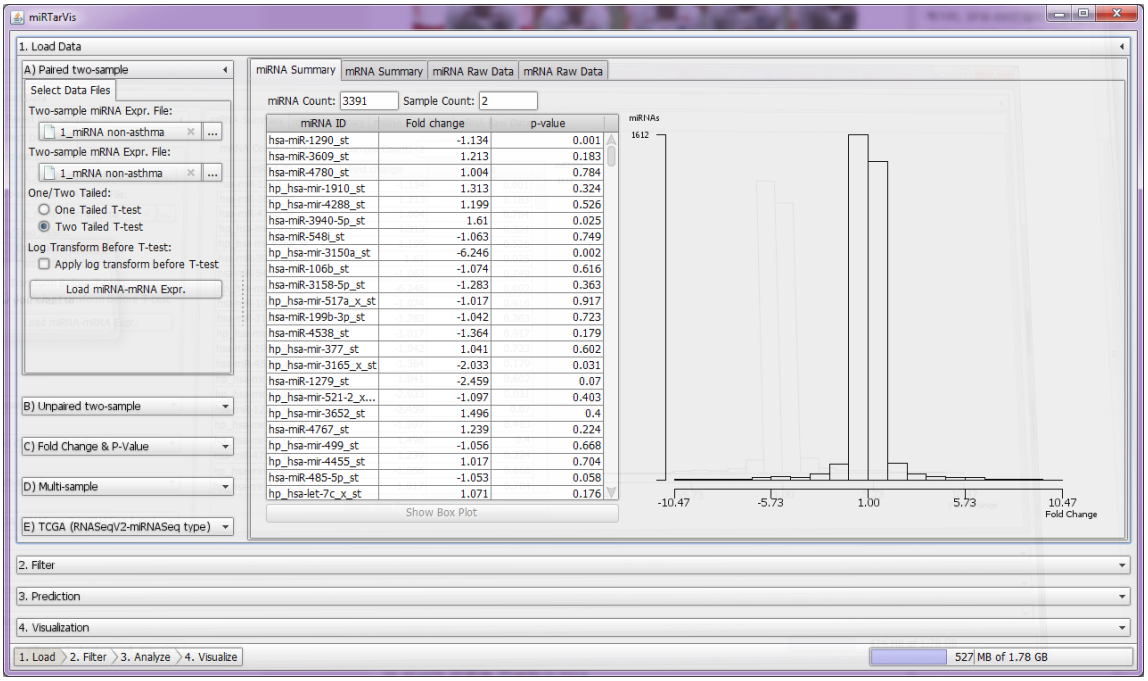
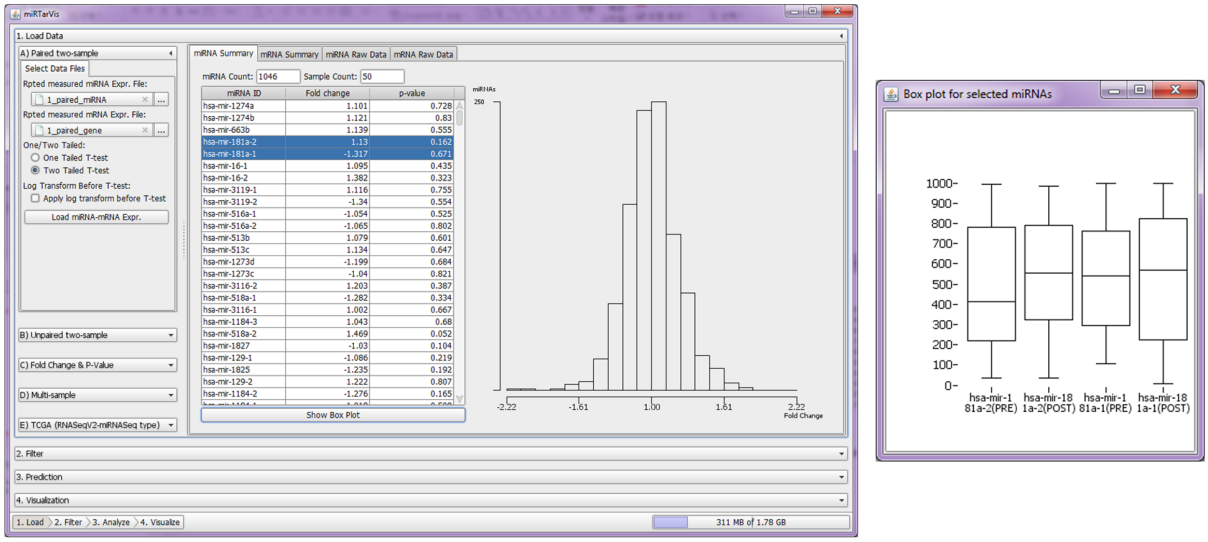



Figure 3.10. Histogram of loaded miRNA and mRNA expression data shows the distribution of input data at the first step of integrated analysis. This helps users to check the normality and outlier in the data before further analysis.

Our interface has four big menus: load data, filter, predict, and visualize. Users can select the data type in the load data menu. The

interface can load one of five types (paired two-sample, unpaired two-sample, p-value and fold change, multi-sample and TCGA) of miRNA-mRNA expression profile data. The interface to load each data type, placed on the left, varies according to which data type a user selects to load. After loading data, miRTarVis shows the data in the main panel on the right. For two-sample miRNA-mRNA expression profile data, the raw data and a summary is shown to users. In the summary, the interface shows automatically calculated fold change and p-value for each miRNA and mRNA, and it shows the distribution of fold change in a histogram (Figure 3.11A). For fold change and p-value type data which does not have underlying data, the interface shows only the summary and histogram of fold change values. For multi-sample data, the interface shows raw miRNA-mRNA expression values and the average and standard deviation of expression levels for each miRNA and mRNA in a table. In addition, it shows a histogram of average expression level. For both two-sample and multi-sample data, box plots can be shown for selected miRNAs and mRNAs (Figure 3.11B).




A) miRTarVis shows histogram of miRNA fold change after loading data

B) Box plots for selected miRNAs

Figure 3.11. Histogram shows the overall distribution of miRNA and mRNA expression. When user select and double click interesting miRNAs and mRNAs, simple box plots pop up, and users can check the distribution for selected miRNA and mRNAs.

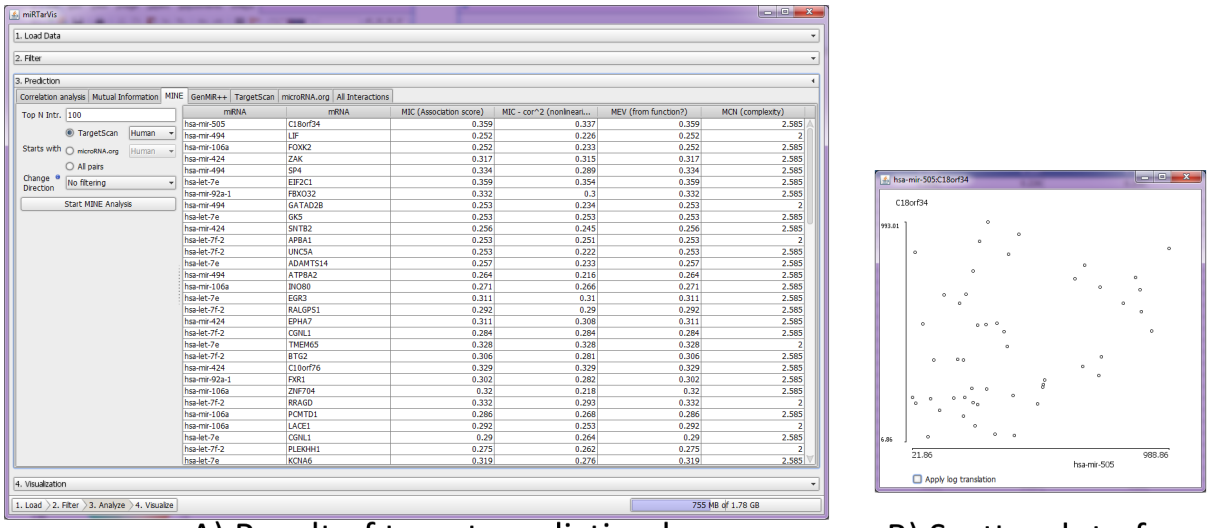



Figure 3.12. The interface deisng for prediction procedure. Users can select prediction option according their own analysis needs. In figure A), users adjusting the parameters for prediction of miRNA-mRNA interactions by MINE analysis technique. Users can start the prediction by pressing the start button. After the prediction execution ends, all results are shown in the table. Users can easily see whether the prediction miRNA-mRNA interaction is supported by their input expression data as the interfaces provides scatterplot of miRNA-mNRA when users double click a row in the table.

In the filter menu, we designed the interface so that users can select only significant miRNAs and mRNAs. The filtered significant miRNAs and mRNAs will serve as a search space in the next predict procedure. For two-sample data type, significant miRNAs and mRNAs are differentially expressed miRNAs and mRNAs. Therefore, users can identify differentially expressed miRNAs and mRNAs by filtering out

miRNAs and mRNAs whose p-values are over a certain threshold or whose fold changes are under a certain threshold.

For multi-sample data type, it is required to make the miRNAs and mRNAs with too low average expression levels excluded in the next predict step. For example, by excluding those miRNAs or mRNAs whose expression levels are zero in almost every sample, the prediction quality is improved because such miRNAs or mRNAs can distort the prediction process. As a result, in the interface, we provides different filtering options for two-sample and multi-sample type data: filtering by p-value or fold-change for the former and filtering by average expression level for the latter. Our interface design shows a histogram of fold change for two-sample type data and a histogram of average expression level for multi-sample type data.

In the predict menu (Figure 3.12), the interface searches for targets of miRNAs by multiple algorithms. The predict menu uses a tab interface where each tab is dedicated to a prediction algorithm. When a user selects a certain prediction algorithm in a tab, interfaces for choosing parameters for the selected algorithm appears on the left. The interface

searches for predicted miRNA-mRNA interactions among all possible miRNAs and mRNAs pairs that are remained from the previous filter step. After finishing the prediction process, the interface presents the predicted miRNAs and their targets in a table on the right. The interface can show a scatterplot of the expression levels of the predicted miRNA-mRNA interaction (Figure 3.12). As this scatterplot showing the expression value relationship between selected miRNA and mRNA directly to users, user can easily see whether the predicted interaction is supported by their data. The miRNAs are known to suppress the expression of their target mRNAs, so if the if miRNA-mRNA interaction is true positive, then the scatter plot have to be seen as negative correlation between miRNA and mRNA.

In the last tab (named "All interactions"), our interface summarizes all the miRNA-mRNA interactions found so far in a table. In the All interactions tab, miRTarVis can intersect prediction results of different user-selected prediction algorithms. This feature is valuable because a miRNA-mRNA interaction commonly predicted by multiple algorithms is more likely to be a genuine miRNA-mRNA interaction.

In the visualize menu, the interface shows the essential part, predicted miRNA-mRNA interactions in node-link diagram or Bipartite Treemap. Node-link diagram in the interface comes in one of four layouts: modified ISOM layout, KK layout, force-directed layout, and circular layout. Modified ISOM and KK layouts are useful to disclose the structure of a predicted miRNA-mRNA regulatory network because they place topologically nearer nodes closer in a physical space.

## 3.4 Comparison with Other Visualization Techniques for MiRNA-mRNA Interaction Network

There are some analysis tool of miRNA-mRNA expression profile data in the Bioinformatics field, and some of them provides visualization for predicted regulatory miRNA-mRNA interaction network. MAGIA [29] and miRConnX [28] are tools that provide visualization for miRNA-mRNA interaction network. We will discuss about their characteristics and compare them with our suggested visualization techniques, Bipartite Treemap and Enhanced Node-link diagram, as denoting their good and bad points.

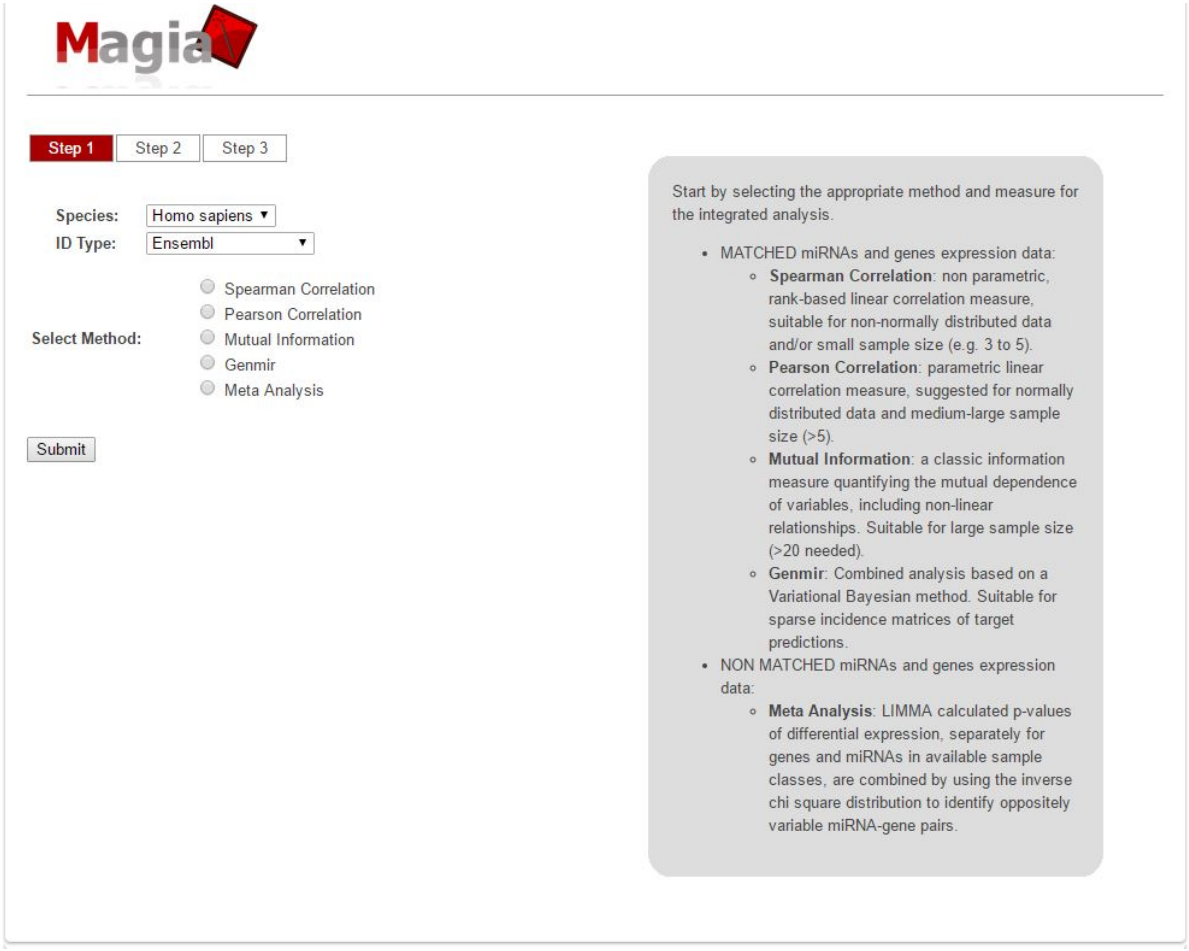



Figure 3.13. The interface of the first step of Magia. As this figure shows, there are three steps in Magia analysis for miRNA-mRNA. In the first step, a user have to select the species, ID type, and method for miRNA-mRNA expression profile based prediction algorithm. As this figure shows, only one miRNA-mRNA expression profile based prediction algorithm can be applied to the input data.

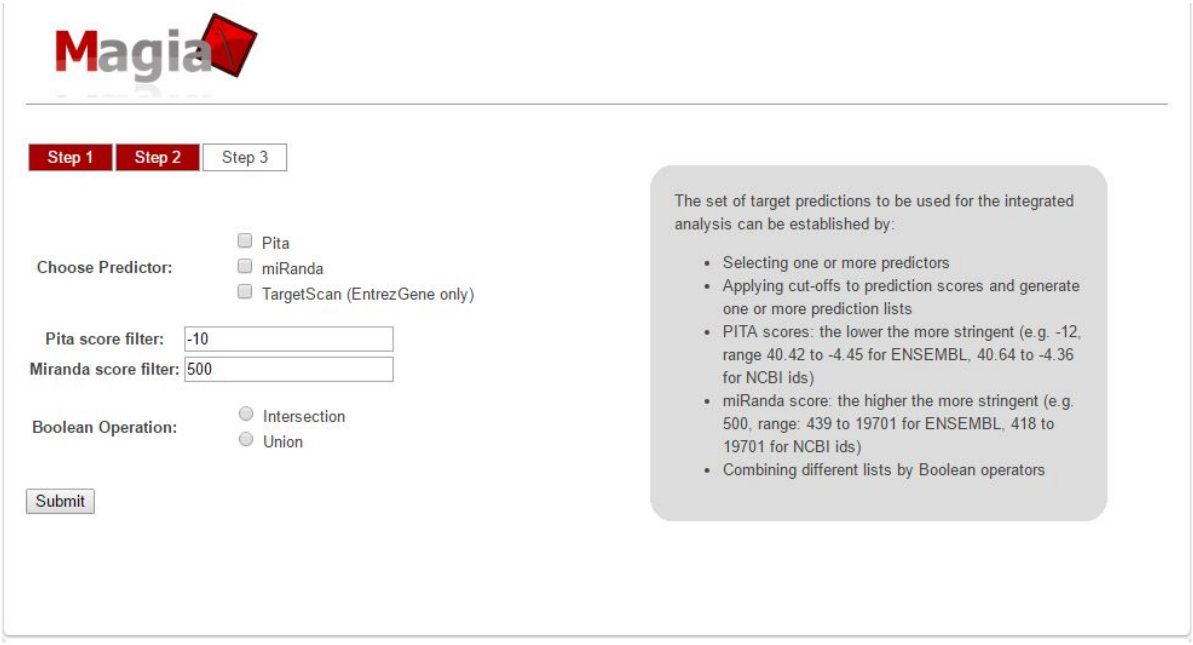



Figure 3.14. This figure show the interface for the second step for Magia analysis of miRNA-mRNA expression data. In this step, users can select multiple sequence prediction algorithms among Pita, miRanda, and TargetScan. In addition, users can select Pita score filter and miRanda score filter cutoff value. Users can select option for how to combine the sequence based prediction algorithms between intersection and union.

MAGIA [29] is a web tool for miRNA-mRNA expression profile data. Figure 3.13 shows the initial step of Magia analysis. Our interface has four procedures for analyzing miRNA-mRNA expression profile data, and Magia has three steps for analysis of miRNA-mRNA expression profile data. In Magia's step 1, a user have to select the species (in Figure 3.13, the species is set as Homo sapiens) and ID type of gene. The ID type of gene represent in what type the gene ids are represents. There are some

methods to denote ids for genes such as EntrezGene type or Ensembl type. The most important parameter that users have to select is the "Select Method" parameter, which represents the miRNA-mRNA expression profile based prediction algorithm. Magia supports only Spearman Correlation, Pearson Correlation, Mutual Information, Genmir, or Meta analysis. One of the limitation of this interface is that only one miRNA-mRNA expression profile based prediction algorithm can be applied to the input data.

In Magia's step 2, a user have to select the sequence prediction algorithms he or she wants to use with their input data. Though the interface allows users to select only one expression profile based prediction algorithm, the interface allows users to select multiple sequence based prediction algorithms.

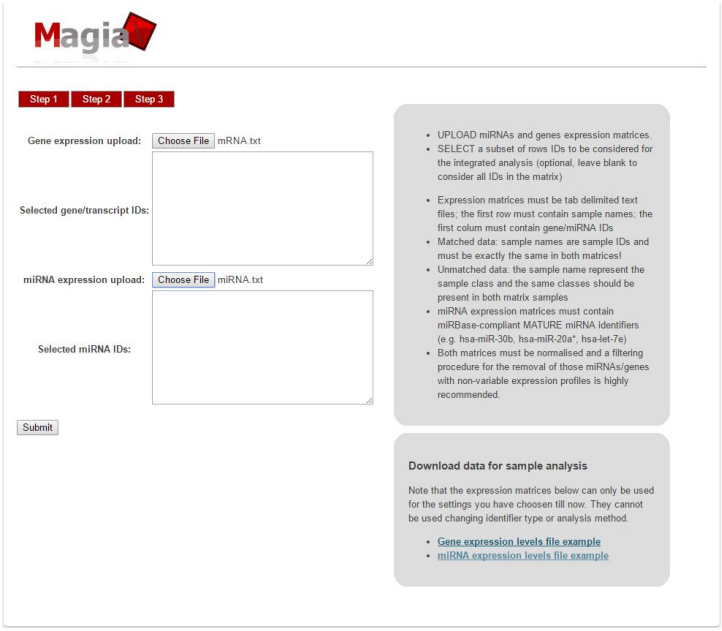



Figure 3.15. This figure shows the third step for Magia analysis for miRNA-mRNA expression profile data. In this step, a user has to specify the miRNA expression profile data file and gene expression profile data file. In the text field, a user can specify the miRNA or gene ids that he or she wants to confine the analysis to the specified miRNAs or genes. If the text field is left as empty, then all miRNAs and genes that are in the input file is analyzed.

In step 3, the interface allows users have to specify the miRNA and mRNA expression profile data files. Figure 3.15 shows the step 3. For the expression profile based prediction, the miRNA expression profile data file and gene expression profile data file have the same sample names. Users can confine the miRNAs and genes that will be considered in the analysis by specifying the id of miRNAs and genes in the text fields. If the text fields are left empty, then all miRNAs and genes in the

input data file are considered in the analysis for predicting miRNA targets.


Figure 3.16. This figure shows the waiting message after a user clicking the submitting button in Magia's analysis step 3.

When a user clicks the submit button, then the waiting interface shows a message that the data is under progress (Figure 3.16). After Magia system finishes the analysis, the results are represented as visualization and table (Figure 3.17). In the node-link diagram, triangle-shaped nodes represent miRNAs and circle-shaped nodes represent genes (mRNAs). The background is black and white links represent predicted miRNA-mRNA interactions. The table in below the visualization lists up all predicted miRNA-mRNA interactions from the previous analysis.

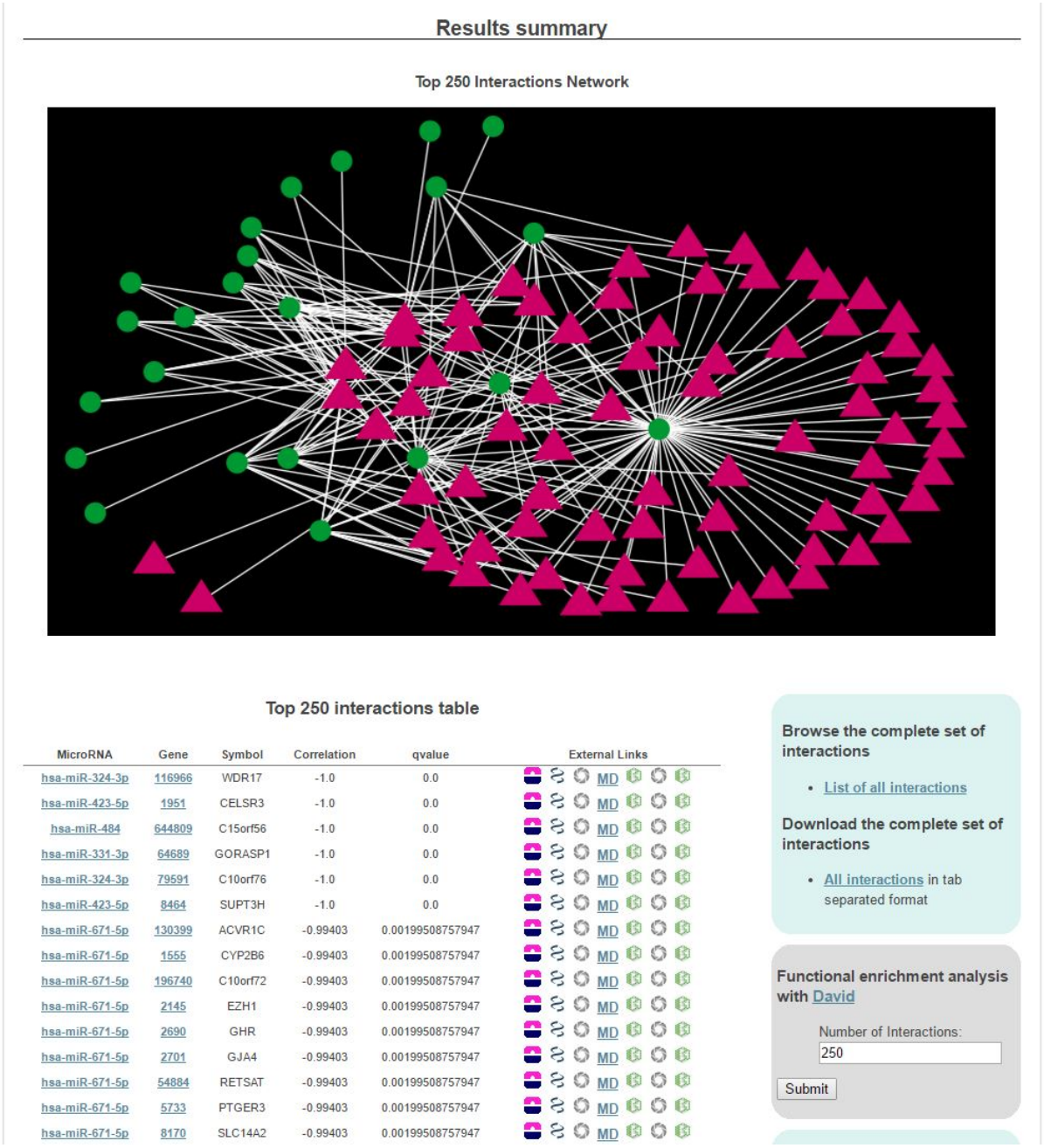



Figure 3.17. Magia present the analysis results in a node-link diagram visualization and table of miRNA-mRNA interactions. In the node-link diagram, the red triangles represent miRNAs, and green circles represent genes (mRNAs). MiRNA-mRNA interactions are represented as solid white lines. miRNA-mRNA interactions are also represented in the table below.

The weakness of the visualization is that it supports very limited interactions. The node-link diagram does not show the labels of nodes. Only when a user hovers cursor over interesting nodes, a small popup show the labels of it. When a user clicks a miRNA node in the node-link diagram, it shows a web-page that shows a table that includes all mRNAs that the miRNA regulates. When a user clicks a gene node, it shows a web-page that shows a table that includes all miRNAs that regulate the mRNA.

Also, Magia's node-link diagram does not show the fold change direction. Since miRNAs down-regulate their target mRNAs, the direction of fold change between miRNA and its predicted target mRNA is an important clue for checking the validation of predicted miRNA-mRNA interaction. However, a user cannot check the fold change information in Magia's node-link diagram.

The graph layout algorithm for Magia's node-link diagram does not effectively show the overall structure of miRNA-mRNA regulatory network. As shown in Figure 3.17, all links are congested and occluded each other like a hair ball. It is not clearly specified which graph layout

algorithm Magia uses the layout algorithm is not suitable for visualizing miRNA-mRNA regulatory network. Furthremore, Magia's node-link diagram edges are solid white. This is another reason for the occlusion problem.

Another problem of Magia's interface is it is not possible to change parameters after the analysis is finished. Once the visualization is created by the tool, and a user wants to change a parameter to create another visualization with different options and the same input data, then the user has to start the procedure again from the start. This makes it difficult to explore the miRNA-mRNA expression profile data in Magia.


Figure 3.18. This figure shows miRConnX's current website. Unfortunately, it is currently out of service.

miRConnX [28] is a tool for analyzing miRNA-mRNA expression profile data. miRConnX is originally available from

http://www.benoslab.pitt.edu/mirconnx/, but the service is currently not available (Figure 3.14). Therefore, instead of showing an example by directly a sample data to miRConnX interface, we use images of miRConnX' s interface from miRConnX paper and manual, and other papers that used and cited miRConnX to discuss its interface design and visualization.

The detail of interface design of miRConnX is not described fully in the miRConnX paper, and it is not currently available, we cannot discuss about miRConnX' s interface design here. We only discuss about the visualizations that show the analysis results by miRConnX.

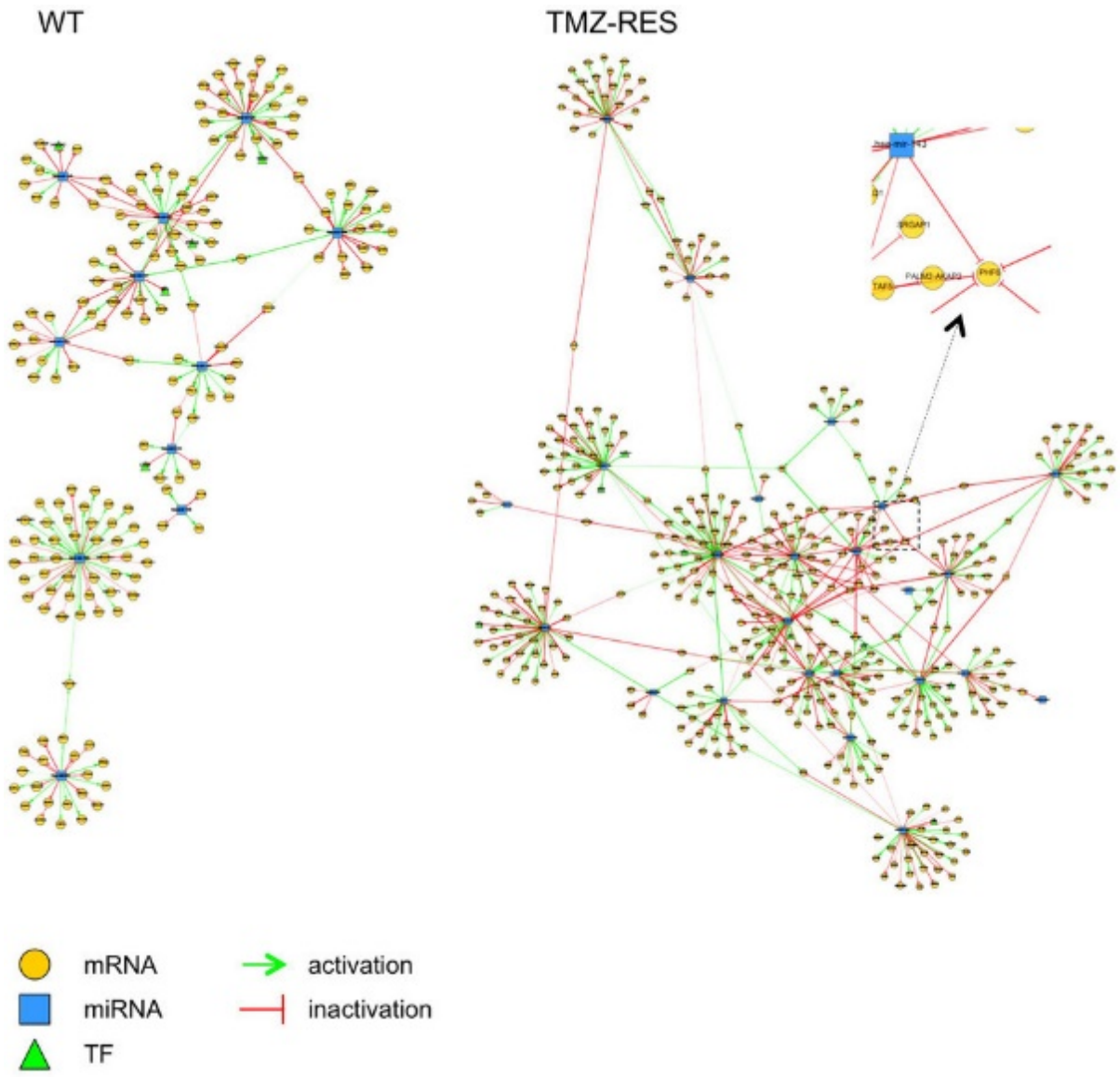



Figure 3.19. This figure shows the visualization generated by miRConnX. The node-link diagram has the better quality than the visualization generated by Magia. Compare this visualization with the visualization in the Figure 3.17. Since miRConnX shows the relationship between TFs and genes, and TFs can up-regulate their target genes, so the node-link diagram has two type of links, activation and inactivation. However, the visualization does not support any direct user interaction in the visualization for analysis of the miRNA-mRNA. Only zoom in/out and panning interaction is supported.

Visualization of miRConnX is generated by Cytoscape. The layout algorithm of mirConnX is much better than that of Magia. Figure 3.19 shows an example figure that generated by miRConnX from Hiddingh et al. [101]. These two node-link diagrams shows the miRNA-mRNA-TF (Transcription Factor) regulatory networks from two miRNA-mRNA expression profile data set respectively. Yellow circle-shaped nodes represents mRNAs, and blue rectangle-shaped nodes represents miRNAs. Green triangle-shaped nodes represents TFs. The type of node is encoded by both color and shape.

However, fold change is not encoded by any visual attributes, so a user cannot see the fold change direction directly in the visualization. An edge is either red or green. Red links represent the inactivation regulatory relationships, and green links represent activation regulatory relationships. However, the link cannot represent any further information such as the possibility of the predicted relationship. Furthermore, there is still possibility of occlusion problem when the number of nodes and links becomes high since the links are represented solid line.

In summary, interfaces and visualizations of Magia and miRConnX are limited when they are compared with our visualizations and interfaces. Magia provides only limited parameters in their analysis, and their visualizations support no interactions. miRConnX is better than Magia at showing the overall structure of miRNA-mRNA interaction network. However, miRConnX supports only simple interactions such as zooming in/out or panning. Our visualization and interface have strong points for analyzing miRNA-mRNA expression profile data when compared with Magia and miRConnX.

# Chapter 4

# miRTarVis

In this section, we will introduce miRTarVis, an interactive visual analysis tool for miRNA-mRNA expression profile data. miRTarVis predicts miRNAs targets from miRNA-mRNA expression profile data and visualizes miRNA-mRNA interaction network that is derived from the prediction. We design the interface of miRTarVis in accordance with the analysis procedure of miRNA-mRNA expression profile data. miRTarVis provides both sequence-based prediction algorithms and expression profile based prediction algorithms. miRTarVis is a visual analysis tool that applies GenMiR++ [20], an advanced miRNA target prediction algorithm that uses a Bayesian inference model. miRTarVis the first tool that introduces Maximal Information-based Nonparametric Exploration (MINE) analysis [71], a new technique which finds highly associated

pairs from multi-dimensional data, to predict targets of miRNAs from miRNA-mRNA expression profile data. miRTarVis can visualize a resulting miRNA-mRNA regulatory network in Bipartite Treemap and enhanced node-link diagram. The Bipartite Treemap is a unique feature of miRTarVis, which is expected to outperform a node-link diagram visualization when miRNA-mRNA interactions are overcrowded. In a case study, we prove the efficacy of miRTarVis by applying it to human miRNA-mRNA expression profile data.

## 4.1 Design goals and Rationale

In the view of visual analysis, we can define the problem of searching for miRNA targets from miRNA-mRNA expression profile data as follows. Input data includes two sets of multi-dimensional data whose dimensionalities (sample count) are equal. Count of elements of two sets equals count of miRNAs and mRNAs respectively. Visual analysis goals are

- to predict miRNA-mRNA interactions from miRNA-mRNA expression profile and

- to understand the structure of the network that consists of the miRNA-mRNA interactions.

We could derive an analysis task for achieving these goals with collaboration with miRNA researchers. The visual analysis task is procedure that consists of four steps:

1) load miRNA-mRNA expression profile data;
2) filter miRNAs-mRNA expression profile data to find and remain only significant miRNAs and mRNAs;
3) predict miRNA-mRNA interaction by searching for highly associated miRNA-mRNA interactions by bioinformatic techniques;
4) visualize the resulting miRNA-mRNA interaction network to help researchers understand the structure of the graph.

*Predict* step is an essential step of the procedure. There are two types of prediction algorithms. Sequence-based target prediction algorithms would give us prior knowledge, so we can regard it as an external static database that gives preset associations between miRNAs and mRNAs.

MiRNA-mRNA expression profile based prediction searches for highly probable associations between miRNAs and mRNAs according to input miRNA-mRNA expression profile data. The analysis result would be a bipartite graph that consists of miRNA-mRNA interactions. In visualize step, Information visualization (Infovis) techniques can enhance exploration and understanding of the network structure.

At the first stage of our iterative design process, we set the two high-level goals which correspond to the visual analysis goals for miRNA-mRNA expression profile data: our analysis tool for miRNA-mRNA expression profile should

- improve performance of miRNA target prediction from miRNA-mRNA expression profile data by integrating multiple target prediction algorithms;
- help comprehension of the resulting miRNA-mRNA interaction network by visualizations.

A long term collaborative design process with biomedical researchers lead us to the following specific design goals to achieve the high-level goals in miRTarVis: It should

- help researchers search for possible miRNAs targets specific to their miRNA-mRNA expression profile data;
- provide a user interface in accordance with the load-filter-predict-visualize procedure for miRNA-mRNA expression profile data;
- combine diverse prediction algorithms and adopt novel prediction algorithms;
- present results in intuitive visualizations;
- support dynamic queries though intuitive user interactions;
- help researchers access information of miRNAs or mRNAs conveniently.

Interfaces of miRTarVis are in accordance with the visual analysis procedure for miRNA-mRNA expression profile data:

1) miRTarVis loads miRNA-mRNA expression profile data;
2) miRTarVis searches for differentially expressed miRNAs and mRNAs for two-sample expression profile data or highly expressed miRNAs and mRNAs for multi-sample data;

3) miRTarVis predicts probable miRNA-mRNA interactions among remaining miRNAs and mRNAs;

4) miRTarVis visualizes a resulting miRNA-mRNA interaction network. Each step of the procedure has a dedicated menu item in miRTarVis.

## 4.2 Input Data

miRTarVis is designed to provide users with better flexibility in data input than previous tools. miRTarVis can accept both two-conditional data (which we can execute differential analysis) and multi-sample miRNA-mRNA expression profile data (differential analysis is not possible for this data). In addition, it accepts data that only consists of fold-change and p-value without underlying expression data. Furthermore, it also directly accepts TCGA miRNA-mRNA expression profile data, and execute differential analysis automatically. We will discuss about them in following paragraphs in detail.

We categorize miRNA-mRNA expression profile data into four types. The first type is paired two-sample (repeated measured) expression

profile data, in which expression levels of every miRNA and mRNA are measured for two paired samples (control vs. treatment; or pre vs. post). For this type of data, miRTarVis performs one tailed or two tailed paired t-test and calculates p-value and fold change for each miRNA and mRNA. miRTarVis uses geometric mean to calculate the fold change:

$$fold\ change = \frac{\sqrt[n]{\prod_n (post)_i}}{\sqrt[n]{\prod_n (pre)_i}},\ n = sample\ size,\ i = sample\ index$$

Furthermore, to process NGS data, we adopt DESeq [93] and EdgeR [102] R packages to calculate fold changes and p-value automatically from reading count data.

The second data type is unpaired two-sample expression profile data. In this type of data, expression levels of every miRNA and mRNA are measured for each sample, and each sample is categorized as either control or treatment (or pre vs. post). miRTarVis performs unpaired t-test under the assumption of either equal or unequal variance, calculating p-value or fold-change for each miRNA and mRNA. For read count data, miRTarVis performs DESeq [93] or EdgeR [102] analysis.

The third data type is p-value and fold change data, which has only p-value and fold change value for every miRNA and mRNA without underlying expression levels. For this type of data, predictions based on expression profile data are not available.

The fourth data type is multi-sample miRNA-mRNA expression profile data, in which expression level of each miRNA and mRNA is measured from multiple samples, but each sample is classified neither as control nor treatment, so it is impossible to conduct t-test. For this data type, miRTarVis just calculates average and standard deviation of expression level for each miRNA and mRNA automatically. In addition, miRTarVis can take expression profile downloaded from TCGA (The Cancer Genome Atlas) data portal as input. It considers the TCGA downloaded data as the multi-sample miRNA-mRNA expression profile data (i.e. the fourth type).

miRTarVis uses a CSV formatted text format. It uses miRBase [64] ID and gene symbol as miRNA and mRNA ID, respectively.

## 4.3 MiRNA Target Prediction and Analysis Procedure

miRTarVis is the first visual analysis tool that adopts the Maximal Information-based Nonparametric Exploration (MINE) analysis [71] and Bayesian inference modeling analysis (GenMiR++ [20]) to search for possible targets of miRNAs from miRNA-mRNA expression profile data. GenMiR++ was based on a Bayesian model to predict targets of miRNAs. This method can search for causal relationships between miRNAs and their targets from miRNA-mRNA expression profile data. However, it is demanding to apply GenMiR++ to particular miRNA-mRNA expression profile data because GenMiR++ algorithm is only available as a Matlab code. miRTarVis helps researchers conveniently run GenMiR++ [20] with their data. The MINE analysis is an algorithm to search for highly associated variable pairs in multi-dimensional data. However, as far as we know, miRTarVis is the first tool to use the algorithm to search for miRNAs targets from miRNA-mRNA expression profile data. miRTarVis also supports correlation analysis and mutual information analysis.

Analysis procedure of miRTarVis consists of load, filter, predict, and visualize. miRTarVis searches for interesting miRNAs and mRNAs in the

filter step. For two-sample expression profile data, those differentially expressed miRNAs and mRNAs are significant, so miRTarVis removes those miRNAs and mRNAs whose p-value is over a user-specified threshold. For multi-sample expression profile data, users can remove miRNAs and mRNAs with very low average expression levels and use remaining miRNAs and mRNAs in the further analysis. This could be useful because such low average expression levels could indicate measurement errors.

In the next *predict* step, miRTarVis searches for miRNA-mRNA target interactions among all remaining miRNA-mRNA pairs from the previous step. miRTarVis integrates both sequence-based prediction algorithms and expression profile based prediction algorithms. miRTarVis originally supports two sequence-based prediction algorithms, TargetScan [62] and microRNA.org [59], which are two of the most cited miRNA target prediction algorithms [40]. Latest version of miRTarVis support five sequence-based prediction algorithms (TargetScan, microRNA.org, MicroCosm [103], PicTar [104], miRDB [105]) by using multiMir R package [106].

miRTarVis supports four expression profile based prediction algorithms, including correlation analysis (a user can choose to use Pearson or Spearman coefficient), mutual information analysis, Bayesian inference model analysis (GenMiR++ [20]), and MINE analysis [71]. Because expression profile based prediction algorithms need expression level at each sample, miRTarVis does not support this type of prediction for the third data type of p-value and fold change data. As discussed earlier, miRTarVis is the first tool that can apply MINE analysis.

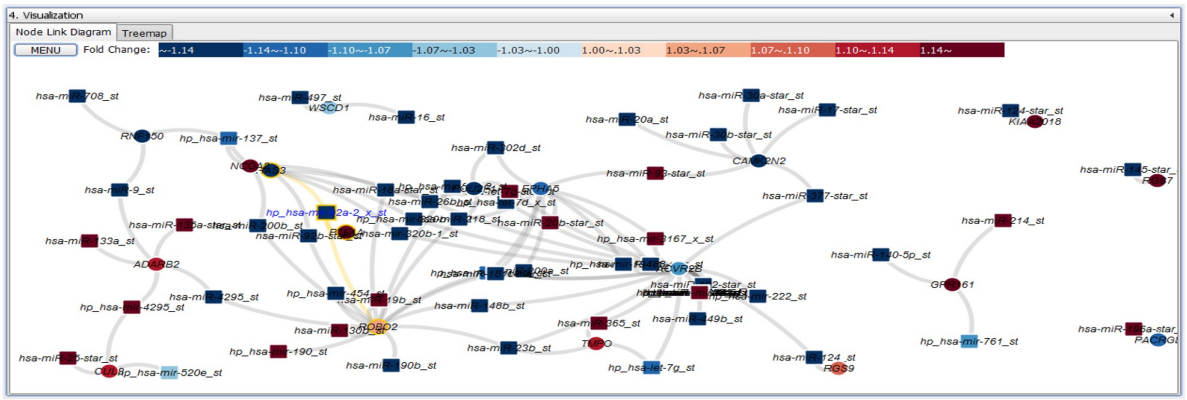
Expression profile based expression profile based prediction algorithms calculate scores representing the intensity of association for each miRNA-mRNA interaction. miRTarVis allows users to set a threshold value to filter miRNA-mRNA interactions by their score when conducting the predictions. In addition, miRTarVis can set the number of resulting miRNA-mRNA interactions for prediction. In this case, miRTarVis searches for the top high-scored miRNA-mRNA interactions.

miRTarVis can filter miRNA-mRNA interactions by their fold change direction as well. For example, predicted miRNA-mRNA interactions where miRNA is up-regulated and mRNA is down-regulated are

biologically more significant because miRNAs down-regulate their target mRNAs. To enable diverse filtering by fold change direction, miRTarVis provides four search options for fold change direction of miRNA-mRNA interactions in prediction: up-regulated miRNA and down-regulated mRNA, down-regulated miRNA and up-regulated mRNA, and oppositely regulated (union of the first and the second option), and all pairs. If prediction is confined to miRNA-mRNA pairs that consist of up-regulated miRNA and down-regulated mRNAs or pairs that consist of down-regulated miRNA and up-regulated mRNAs, the accuracy of prediction could be improved. Therefore, miRTarVis enables users to use these options to improve accuracy of their miRNA target prediction.

In the next visualize step, miRTarVis generates a regulatory network between miRNAs and their targets from the target prediction result in the previous predict step, and visualizes the resulting bipartite graph both in a node-link diagram and a Treemap. In a node-link diagram (Figure 4.1A, Figure 4.2A), miRNAs and mRNAs are represented as rectangular and circular shaped nodes respectively, and miRNA-mRNA interactions are represented by links between them. For two-sample

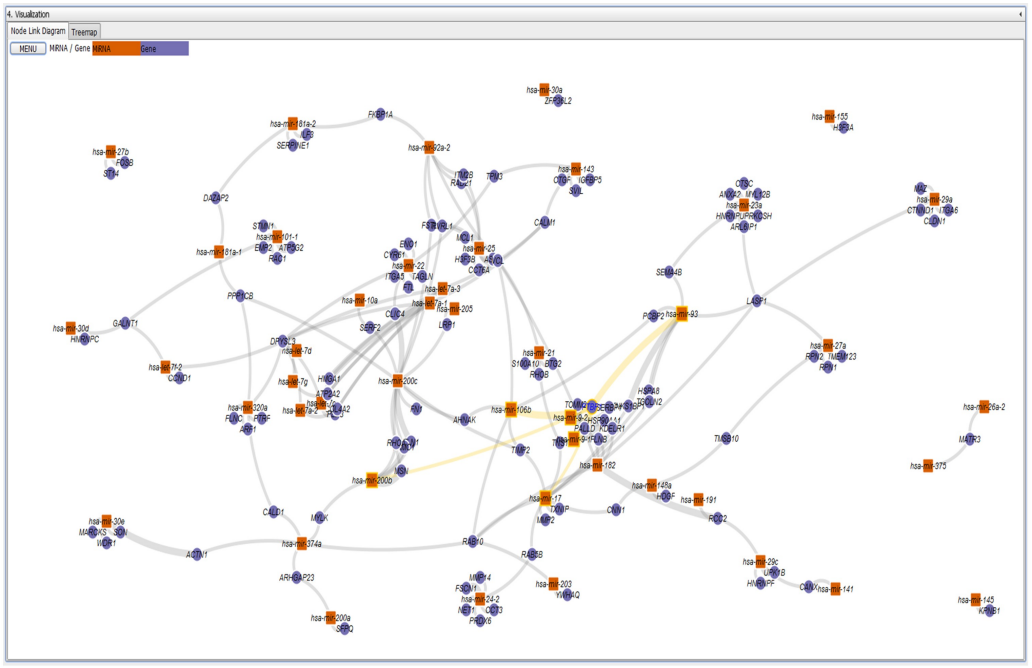
data, fold change value is color-coded: up-regulated miRNAs and mRNAs in red, and down-regulated miRNAs and mRNAs in blue. The intensity of fold change is represented by color saturation while darker color means higher degree of fold change (Figure 4.1).



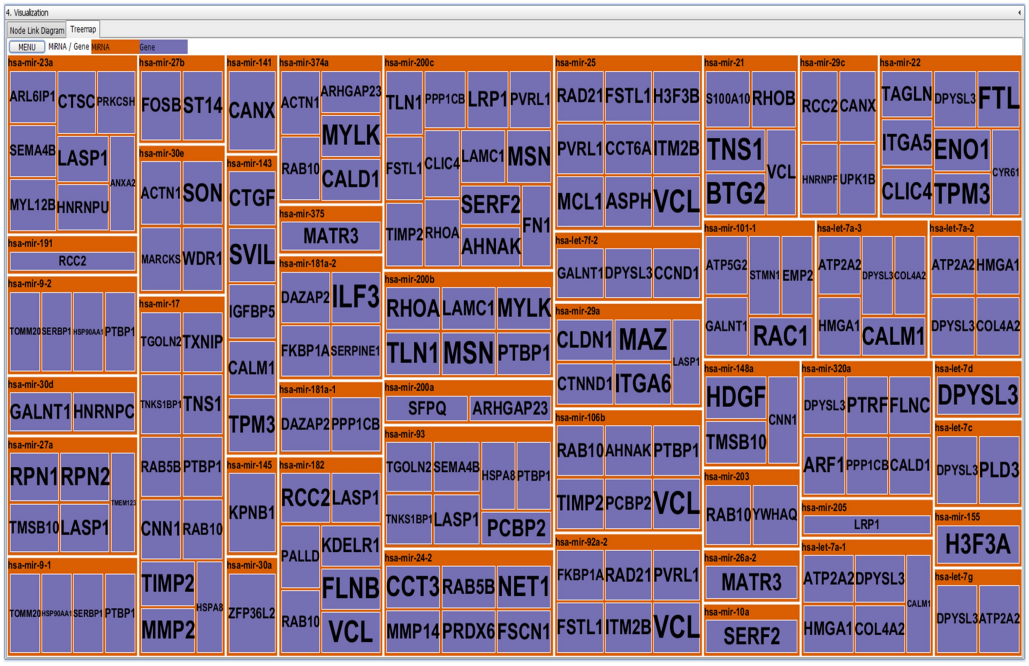

A) Node-link diagram


B) Treemap

Figure 4.1. This figure shows visualization of miRNA-mRNA interaction regulatory network by A) node-link diagram and B) treemap. Red and blue colors represent up- and down-regulated fold changes, respectively. Color saturation represents the intensity of fold changes.



A) Node-link diagram



B) Treemap

Figure 4.2. *miRTarVis* visualizes a miRNA-target interaction networks of miRNA-mRNA expression profiles from 100 TCGA breast cancer samples data in A) node-link diagram and B) treemaps.

This color encoding of fold change is the unique visual cue in miRTarVis to help users easily recognize whether a predicted miRNA-mRNA interaction is supported by input experimental miRNA-mRNA expression profile data. A miRNA down-regulates expression levels of its target mRNAs. Therefore, by comparing the colors of the ends of a link, a user can easily grasp whether the prediction is supported by the input data or not. For multi-sample miRNA-mRNA expression profile data, miRTarVis represents miRNAs in orange and mRNAs in dark blue (Figure 4.2). If some part of a network looks interesting, a user can navigate (zooming and panning) around a specific miRNA with simple mouse wheel interaction in the node-link diagram.

In addition, miRTarVis can visualize a miRNA-mRNA regulatory network using a Treemap (Figure 4.1B). Treemap can represent a complex regulatory network without occlusion among links and nodes especially when there are too many links crossing each other in a node-link diagram. For example, in Figure 4.1, a miRNA-mRNA interaction network is visualized in both Treemap and node-link diagram. In the middle part of the node-link diagram, it is more difficult than Treemap

to identify targets of a miRNA because many nodes are congested. miRTarVis represents all miRNAs as top-level nodes and all their targets as their child nodes. As a result, if multiple miRNAs have common target mRNA, the mRNA node appears multiple times in the Treemap. The color encoding is the same as in the node-link diagram.

In the visualize step, users can efficiently confirm the overview of the resulting miRNA-mRNA regulatory network structure. If the result is not satisfying, *miRTarVis* enables users to go back to *filter* or *predict* steps to change the parameters of filtering or prediction for a better result. Through this iterative procedure, *miRTarVis* can help users efficiently narrow down to more important and interesting miRNA-mRNA interactions.

## 4.4 Visualizations in miRTarVis

miRTarVis is the first tool that adopts a Treemap to show a resulting miRNA-mRNA regulatory network. The Treemap is more effective than the traditional node-link diagram when the network is overcrowded by a large number of miRNAs-target interactions. miRTarVis also

visualizes a miRNA-mRNA network in a more interactive node-link diagram. Users can navigate (i.e. zoom and pan) the diagram for closer inspection of an interesting miRNA or mRNA. In addition, users can move miRNA or mRNA nodes to a better location.

This enhanced interactivity can help users understanding the structure of a miRNA-mRNA network and can polish up the quality of a node-link diagram to be more suitable for publication. *miRTarVis* also helps researchers access external links for miRNAs and mRNAs. If a user wants to see detailed information about a certain miRNA or mRNA, a few mouse clicks open the corresponding web site containing the information. We will explain this in detail when we later describe the design and implementation of miRTarVis.

miRTarVis also provides gene enrichment analysis (provided by the Gene Ontology Consortium web service) of target genes by a miRNA. With accurate prediction, biological functions of predicted targets of a miRNA tend to be similar. By providing a term enrichment analysis, miRTarVis gives a convenient way for confirming the function of a miRNA and the validity of target prediction result.

## 4.5 Implementation

miRTarVis, implemented using Java, runs on any systems with JRE version 1.7 or higher. miRTarVis used the first method of the two methods presented by Kraskov et al. [107] for mutual information estimation. Contrast to conventional mutual information estimators using binning of variables, the method by Kraskov et al. [107] used k-nearest neighbor distances to estimate mutual information, resulting in a better precision. We implemented the algorithm in Java and integrated it into miRTarVis.

The GenMiR++ [20] algorithm, Bayesian inference analysis for miRNA target prediction, is originally implemented in Matlab. We converted it into Java and integrated it into miRTarVis. We validated our implementation by checking whether our result is the same as that of original Matlab implementation using the miRNA-mRNA expression profile data submitted by the authors of GenMiR++ [20] (151 human miRNAs and 16,063 mRNAs across 88 tissues).

The MINE analysis finds highly associated variable pairs in multivariate dataset. It aims at improving generality (not limited to specific function types, i.e. linear or exponential) and equitability (similar score for equally noisy relationships of different types) of analysis. It calculates a maximal information coefficient (MIC) score, which reflects the intensity of association between two variables. We implemented the MINE analysis algorithm [108] in Java and integrated it into miRTarVis.

miRTarVis embedded TargetScan databases. We downloaded a miRNA family table and a predicted conserved target information table, and joined the two tables into one. The resulting table contains three attributes: miRBase ID, gene symbol, and species. miRTarVis has the resulting table that contains a set of conserved targets of all miRNAs for nine species (human, mouse, rat, rhesus, frog, dog, cow, chimpanzee, and chicken). miRTarVis also embedded databases from microRNA.org (www.microrna.org). We downloaded target predictions with good mirSVR scores and conserved miRNA from August 2010 release from the website and embedded it into miRTarVis.

# Chapter 5

# Case Study

## 5.1 Analysis of miRNA-mRNA Expression Profile Data from Asthmatic and Non-asthmatic Cells by miRTarVis

As previous tools for miRNA-mRNA expression data, we applied miRTarVis to analyze miRNA-mRNA expression profile data from an experiment. To give more bioinformatic significant to the analysis, we cooperated with a biologist who generated the data. In the following paragraphs, we will describe basic biological background of the data and report the analysis result by miRTarVis.

As the rate of comorbid asthma and obesity increases, identifying mechanisms by which obesity affects asthma is critical. It is reported that obese visceral adipocytes shed exosomes containing miRNAs that

can up-regulate the expression of profibrotic signaling genes in the lung [109]. An important next step in our analyses was to define the set of lung mRNA responses to these adipocyte-derived exosomes. Prior standard approaches would have included generating a list of potential target mRNAs and prioritizing them for validation. miRTarVis presented us with a new opportunity to objectively define our target validation set of mRNAs using multiple in silico analyses.

Using airway fibroblasts (i.e., cells important in the development of lung fibrosis), we demonstrated the use of miRTarVis to define potential target mRNAs through which obesity can induce lung fibrosis in asthma. We used obese visceral adipocyte-derived exosomes (n = 4) that were previously tested for miRNA expression (Affymetrix microRNA 3.0 array). We coincubated these exosomes with human airway fibroblasts (from endobronchial biopsy tissue) from nonasthmatic and asthmatic donors (n = 1 each) for 24 hours. Fibroblasts were profiled for global mRNA expression.

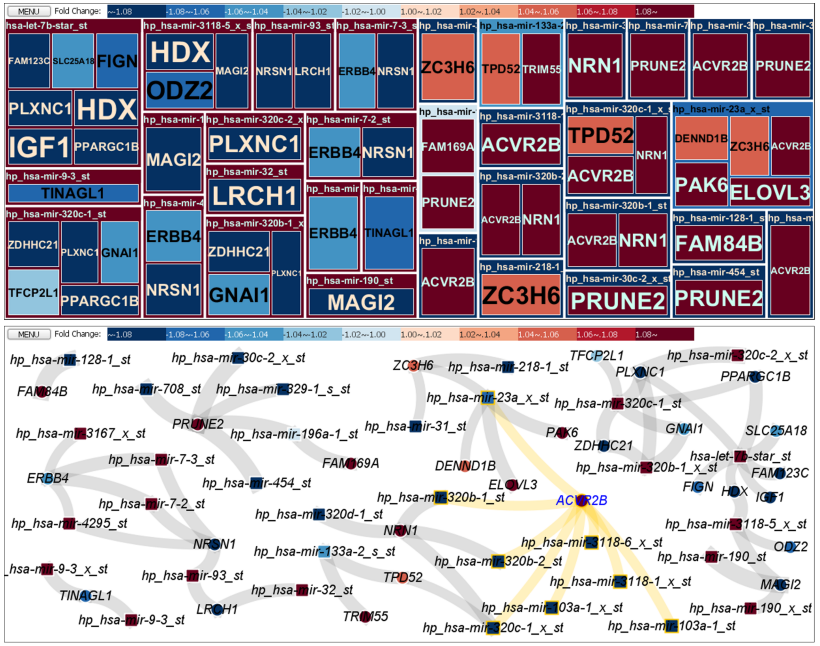
One of the major advantages of miRTarVis is that it enables instantaneous application of multiple analytical algorithms to the data.

Normalized, background-subtracted miRNA-mRNA expression profile data were imported into miRTarVis and filtered for a paired two-tailed t-test with $p \leq 0.05$. Multiple prediction algorithms (i.e., Pearson correlation, MINE, GenMiR++, and TargetScan) were applied, and the top 1,000 negative correlations and top 100 opposite change direction were selected in each algorithm. The intersection among them could be efficiently identified through the filter step of miRTarVis.

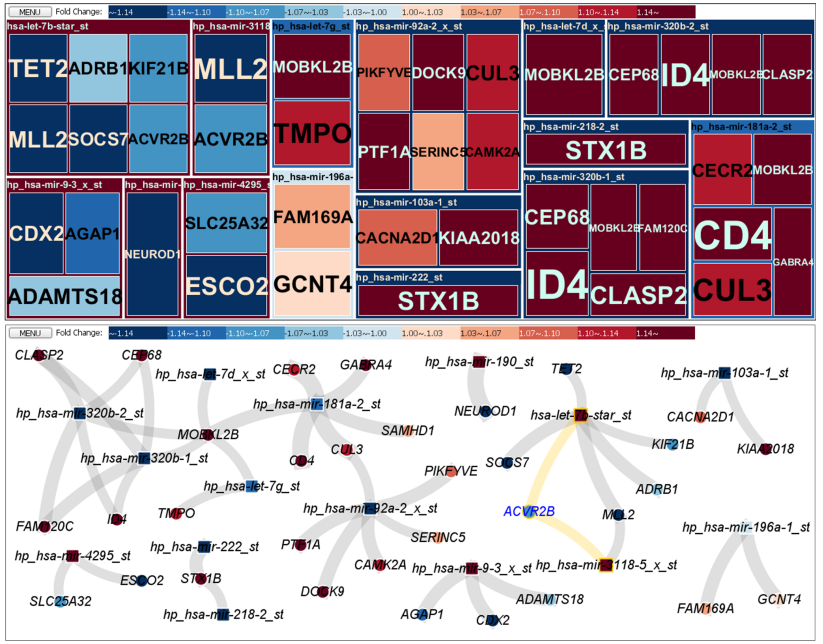
As shown in Table 5.1 and Table 5.2, we identified 61 miRNA-mRNA pairs (33 miRNAs / 27 mRNAs) for asthmatic fibroblasts, and 45 miRNA-mRNA pairs (15 miRNAs / 33 mRNAs) for obese visceral exosomes and nonasthmatic fibroblasts. Our focus turned to ACVR2B (activin receptor, type IIB; myostatin and TGF$\beta$ receptor) as the only gene present in both datasets, that is, down-regulated in nonasthmatic fibroblasts (fold change [FC] = -1.18, $p < 0.01$) and up-regulated in asthmatic fibroblasts (FC = 1.31, $p = 0.02$). Figure 5.1 shows that obese visceral exosomal miRNAs targeting ACVR2B were up-regulated in nonasthmatic fibroblasts (i.e., hsa-let-7b-star_st [FC = 2.31, $p = 0.027$] and hp_hsa-mir-3118-5_x_st [FC = 2.16, $p = 0.025$]) and down-regulated in

asthmatic fibroblasts (hp_hsa-mir-103a-1_st [FC = -2.47, p < 0.001], hp_hsa-mir-103a-1_x_st [FC = -2.18, p = 0.003], hp_hsa-mir-23a_x_st [FC = -1.08, p = 0.035], hp_hsa-mir-3118-1_x_st [FC = -1.53, p = 0.029], hp_hsa-mir-3118-6_x_st [FC = -1.56, p = 0.008], hp_hsa-mir-320b-1_st [FC = -2.41, p = 0.002], hp_hsa-mir-320b-2_st [FC = -1.51, p = 0.019], and hp_hsa-mir-320c-1_x_st [FC = -2.37, p = 0.004]). qRT-PCR confirmed ACVR2B down-regulation in nonasthmatic fibroblasts (FC = 0.26, 95% confidence interval = [0.26, 0.78]) and up-regulation in asthmatic fibroblasts (FC = 3.21, [3.21, 6.72]).

In summary, miRTarVis analyses quickly and inexpensively identified a biologically relevant mRNA target for adipocyte-derived exosomal miRNAs. This target, ACVR2B, is down-regulated in nonasthmatic fibroblasts and up-regulated in asthmatic fibroblasts, suggesting that obese visceral adipocyte-derived exosomes regulate airway fibroblast gene expression and that these cells respond differently to the exosomes depending on disease state. miRTarVis enabled this novel mechanistic discovery by which adiposity may increase lung fibrosis in asthma.



A) miRNA-mRNA interaction network for asthmatic



B) miRNA-mRNA interaction network for non-asthmatic

Figure 5.1. miRTarVis visualizes miRNA-mRNA interaction network from asthmatic and non-asthmatic fibroblasts exposed to obese visceral exosomes. The user could identify that ACVR2B is differentially expressed in both conditions in opposite direction (up-regulated in asthmatic and down-regulated in non-asthmatic). In addition, he could identify list of miRNAs that could regulate ACVR2B.

Table 5.1. miRNA-mRNA interactions for obese visceral exosomes and asthmatic fibroblasts

| microRNA | mRNA | Correlation | GenMiR++ | TargetScan | MINE | miRanda |
|---|---|---|---|---|---|---|
| hp_hsa-mir-3118-1_x_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320b-1_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-103a-1_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-3118-6_x_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320b-2_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320c-1_x_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-23a_x_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-103a-1_x_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-23a_x_st | DENND1B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-23a_x_st | ELOVL3 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-4295_st | ERBB4 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-3167_x_st | ERBB4 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-7-2_st | ERBB4 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-7-3_st | ERBB4 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hsa-let-7b-star_st | FAM123C | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-196a-1_st | FAM169A | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-128-1_st | FAM84B | TRUE | TRUE | TRUE | FALSE | TRUE |
| hsa-let-7b-star_st | FIGN | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320c-1_st | GNAI1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320b-1_x_st | GNAI1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-3118-5_x_st | HDX | TRUE | TRUE | TRUE | FALSE | FALSE |
| hsa-let-7b-star_st | HDX | TRUE | TRUE | TRUE | FALSE | FALSE |
| hsa-let-7b-star_st | IGF1 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-32_st | LRCH1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-93_st | LRCH1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-3118-5_x_st | MAGI2 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-190_st | MAGI2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-190_x_st | MAGI2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320d-1_st | NRN1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320b-1_st | NRN1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320b-2_st | NRN1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320c-1_x_st | NRN1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-4295_st | NRSN1 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-93_st | NRSN1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-7-2_st | NRSN1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-7-3_st | NRSN1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-3118-5_x_st | ODZ2 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-23a_x_st | PAK6 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320c-2_x_st | PLXNC1 | TRUE | TRUE | TRUE | TRUE | TRUE |
| hp_hsa-mir-320c-1_st | PLXNC1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320b-1_x_st | PLXNC1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hsa-let-7b-star_st | PLXNC1 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320c-1_st | PPARGC1B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hsa-let-7b-star_st | PPARGC1B | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-30c-2_x_st | PRUNE2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-708_st | PRUNE2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-329-1_s_st | PRUNE2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-454_st | PRUNE2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-196a-1_st | PRUNE2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hsa-let-7b-star_st | SLC25A18 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320c-1_st | TFCP2L1 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-9-3_st | TINAGL1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-9-3_x_st | TINAGL1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-133a-2_s_st | TPD52 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320c-1_x_st | TPD52 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-133a-2_s_st | TRIM55 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-218-1_st | ZC3H6 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-31_st | ZC3H6 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-23a_x_st | ZC3H6 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320c-1_st | ZDHHC21 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320b-1_x_st | ZDHHC21 | TRUE | TRUE | TRUE | FALSE | TRUE |

Table 5.2. miRNA-mRNA interactions for obese visceral exosomes and non-asthmatic

| microRNA | mRNA | Correlation | GenMiR++ | TargetScan | MINE | miRanda |
|---|---|---|---|---|---|---|
| hp_hsa-mir-3118-5_x_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hsa-let-7b-star_st | ACVR2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-9-3_x_st | ADAMTS18 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hsa-let-7b-star_st | ADRB1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-9-3_x_st | AGAP1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-103a-1_st | CACNA2D1 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-92a-2_x_st | CAMK2A | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-181a-2_st | CD4 | TRUE | TRUE | TRUE | TRUE | TRUE |
| hp_hsa-mir-9-3_x_st | CDX2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-181a-2_st | CECR2 | TRUE | TRUE | TRUE | TRUE | TRUE |
| hp_hsa-mir-320b-2_st | CEP68 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320b-1_st | CEP68 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320b-2_st | CLASP2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320b-1_st | CLASP2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-92a-2_x_st | CUL3 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-181a-2_st | CUL3 | TRUE | TRUE | TRUE | TRUE | TRUE |
| hp_hsa-mir-92a-2_x_st | DOCK9 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-4295_st | ESCO2 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320b-2_st | FAM120C | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320b-1_st | FAM120C | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-196a-1_st | FAM169A | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-181a-2_st | GABRA4 | TRUE | TRUE | TRUE | TRUE | FALSE |
| hp_hsa-mir-196a-1_st | GCNT4 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-320b-2_st | ID4 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320b-1_st | ID4 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-103a-1_st | KIAA2018 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hsa-let-7b-star_st | KIF21B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-3118-5_x_st | MLL2 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hsa-let-7b-star_st | MLL2 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-let-7d_x_st | MOBKL2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320b-2_st | MOBKL2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-320b-1_st | MOBKL2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-let-7g_st | MOBKL2B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-181a-2_st | MOBKL2B | TRUE | TRUE | TRUE | TRUE | FALSE |
| hp_hsa-mir-190_st | NEUROD1 | TRUE | TRUE | TRUE | TRUE | TRUE |
| hp_hsa-mir-92a-2_x_st | PIKFYVE | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-92a-2_x_st | PTF1A | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-181a-2_st | SAMHD1 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-92a-2_x_st | SERINC5 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-mir-4295_st | SLC25A32 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hsa-let-7b-star_st | SOCS7 | TRUE | TRUE | TRUE | FALSE | TRUE |
| hp_hsa-mir-222_st | STX1B | TRUE | TRUE | TRUE | TRUE | FALSE |
| hp_hsa-mir-218-2_st | STX1B | TRUE | TRUE | TRUE | FALSE | FALSE |
| hsa-let-7b-star_st | TET2 | TRUE | TRUE | TRUE | FALSE | FALSE |
| hp_hsa-let-7g_st | TMPO | TRUE | TRUE | TRUE | FALSE | FALSE |

## 5.2 Analysis of miRNA-mRNA Expression Profile Data using TCGA Breast Cancer Dataset

To verify the effectiveness of miRTarVis by comparing with other tools, we conduct another case study using a public dataset from TCGA (The Cancer Genome Atlas) by miRTarVis. We downloaded miRNA-mRNA expression profile data from 60 cell lines. Among them, 50 samples are normal cell lines and 10 samples are normal cell lines. We select the miRNASeq and MiRNASeqV2 types for TCGA download data parameter. We will describe the analysis by miRTarVis in detail systematically and report the result.

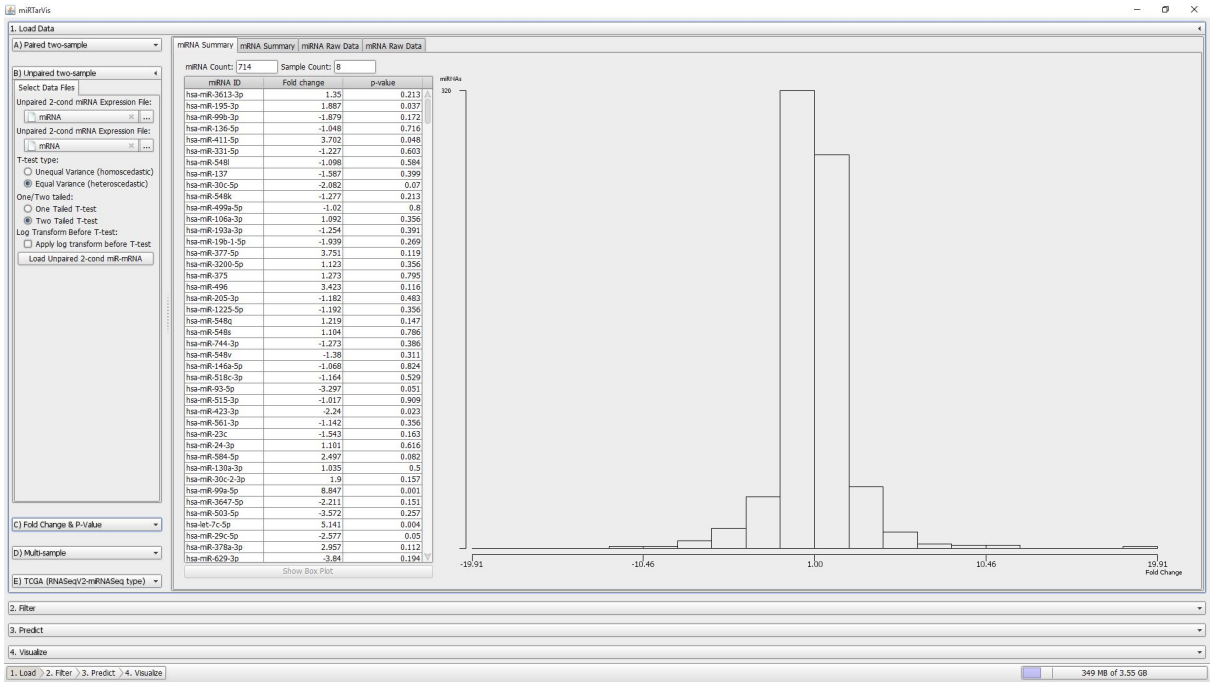



Figure 5.2. The load data step for TCGA data from breast cancer cell lines. The data contains 10 normal cell lines, and 50 cancer cell lines. As a user load a data, miRTarVis shows a histogram that presents the distribution of fold change of the data. For this data, the distribution of miRNA fold change is close to normal distribution.

In the load step of miRTarVis, we set the data type as unpaired two-sample data type, as the number of cancer cells and normal cells are different. We set t-test type as unequal variance since the variances of miRNAs and mRNAs are expected to be different, and we select t-test mode as two-tailed t-test. When a user load input miRNA-mRNA expression profile data, histograms appears to show the distribution of fold change of miRNAs and mRNAs (Figure 5.2).

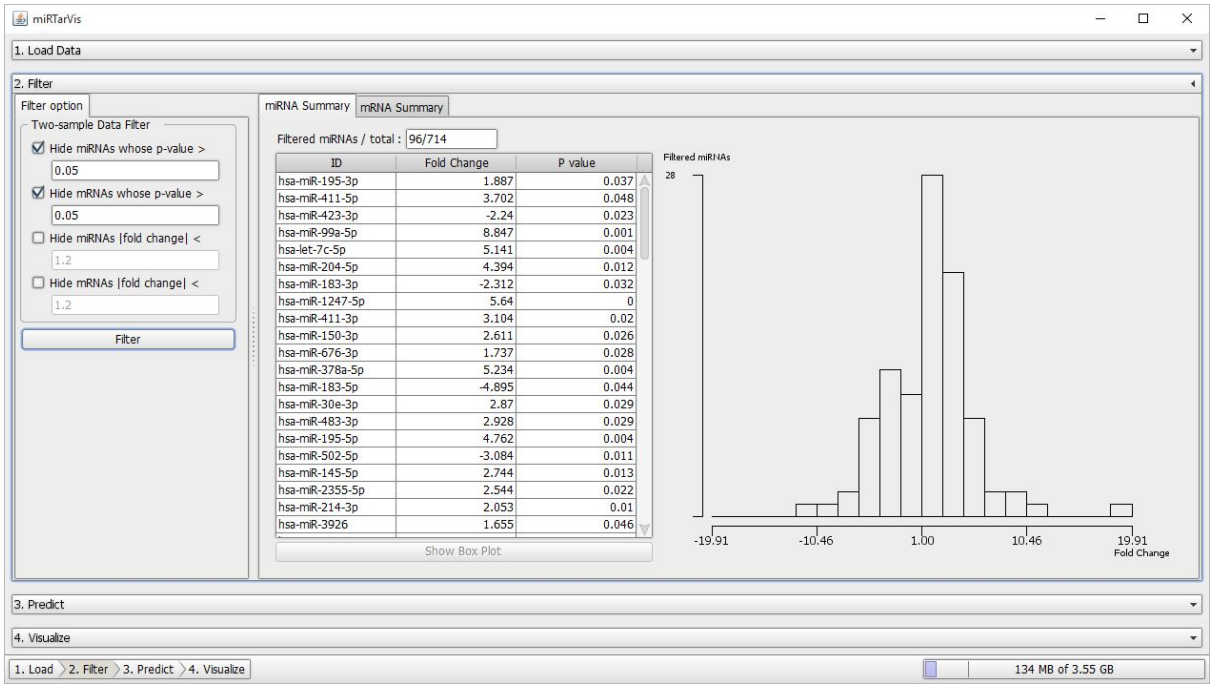



Figure 5.3. The second step of miRTarVis, filter step, a user can filter out insignificant miRNAs and mRNAs for further analysis in step 3. In miRTarVis, users can filter by p-value and fold change. In this figure, the user filter out those miRNAs and mRNAs whose p-value is greater than 0.05.

In filter step, a user can remain only significant miRNAs and mRNAs. Figure 5.3 shows the filtering step in our case study. The remaining miRNAs and mRNAs are involved in the next analysis step. The table shows the remaining miRNAs and mRNAs to the user. The histogram on the right is also updated dynamically as a user changes the filtering parameters.

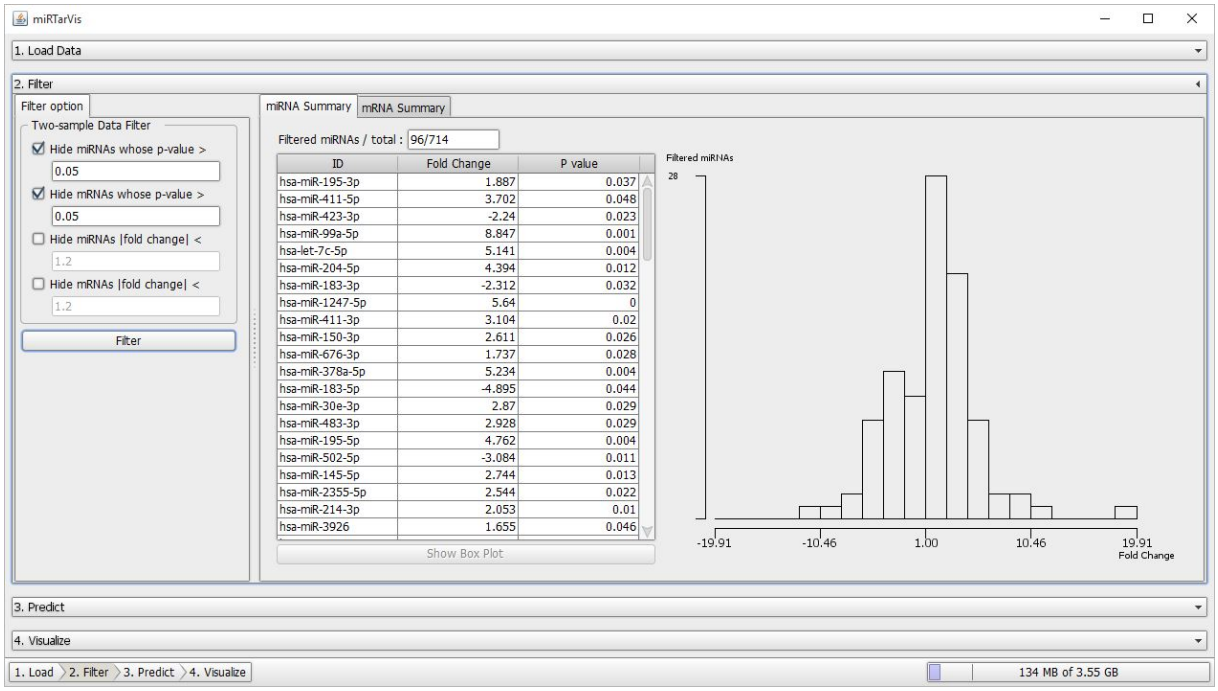



Figure 5.4. This figure shows the prediction step in our case study with TCGA breast cancer data. We remove those miRNAs and mRNAs whose p-value is under 0.05.

In prediction step, we use correlation analysis (Pearson coefficient) with only negative correlation, mutual information, MINE, GenMiR++, TargetScan and miRanda. All the predicted miRNA-mRNA interactions are shown in Table 5.3. In visualization step, the predicted miRNA-mRNA regulatory network are visualized by Bipartite Treemap and enhanced node-link diagram (Figure 5.5 shows the enhanced node-link diagram, and Figure 5.6 shows Bipartite Treemap).

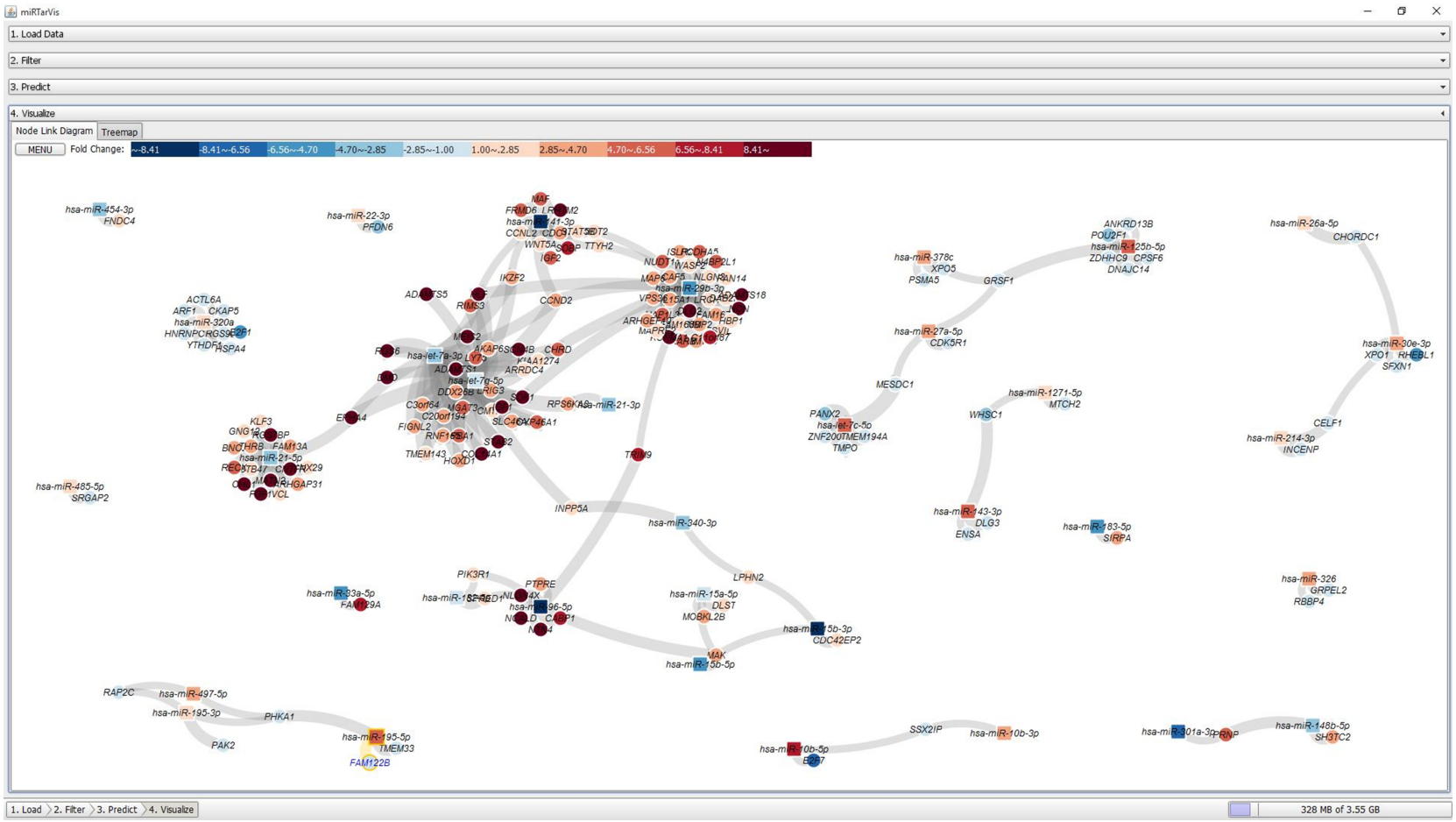



Figure 5.5. This figure shows the enhanced node-link diagram by the miRNA-mRNA expression profile data from TCGA breast cancer dataset. The thickness of links represents how significant the prediction is. In this node-link diagram, CCND2 gene is predicted to be regulated by multiple miRNAs (hsa-let-7a-3p, hsa-let-7a-5p, hsa-miR-29b-3b, and hsa-miR-141-3p).

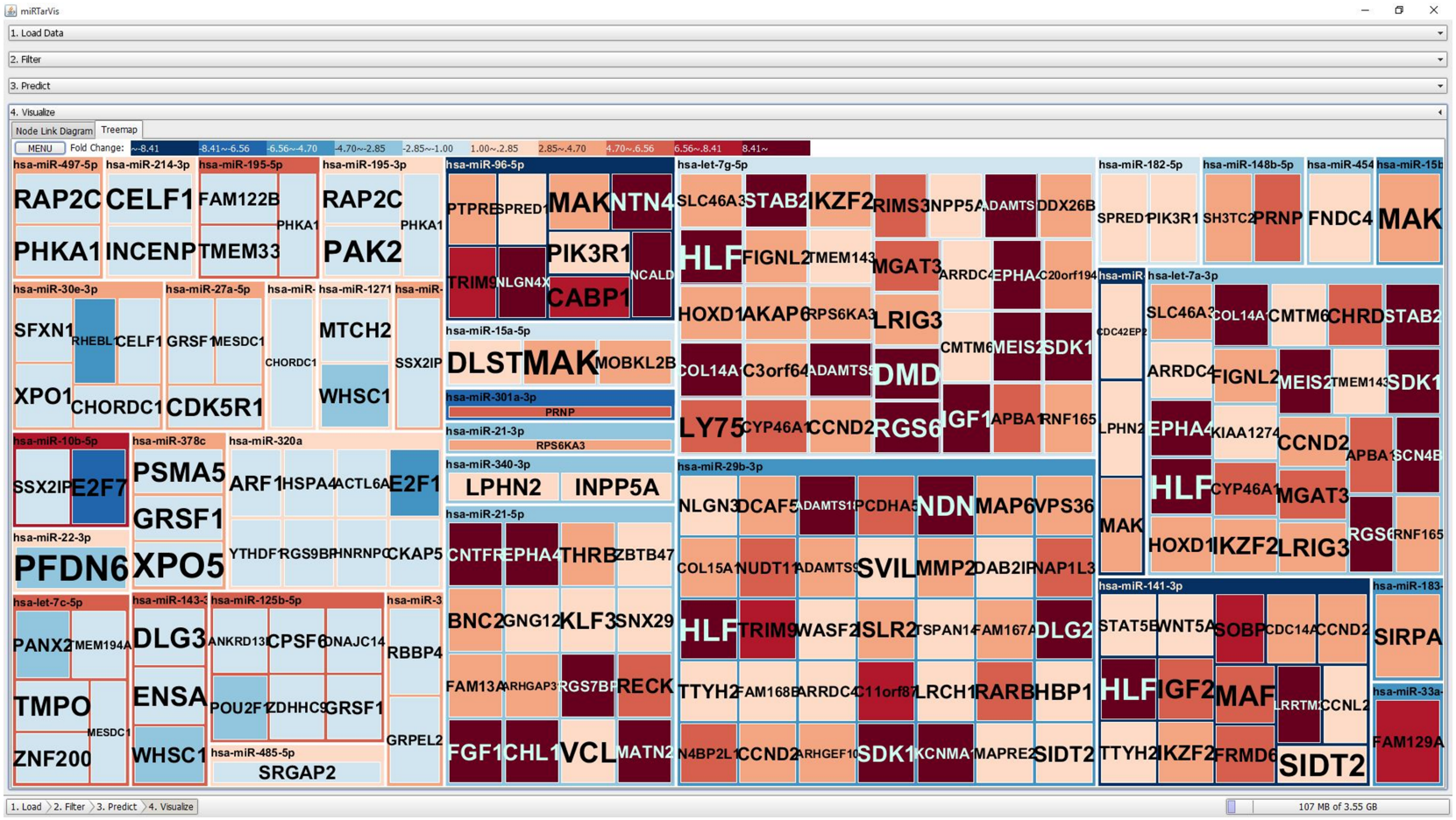



Figure 5.6. This figure shows the Bipartite Treemap for predicted miRNA-mRNA regulatory network in our case study with miRNA-mRNA expression profile data from TCGA breast cancer data. Two miRNAs (hsa-let-7g-5p and hsa-miR-29b-3p are predicted to have specially many target mRNAs. We could expect that these two miRNAs many play an important role in this dataset.

As one can see in the Bipartite Treemap (Figure 5.6), the miRNA that has the most predicted target mRNA was hsa-miR-29b-3p. According to literature ( [110], [111], [112]), miR-29 plays an important role in development of cancer. Furthremore, as one can see in the enhanced node-link diagram (Figure 5.5), CCND2 is an important mRNA that is predicted to be regulated by multiple miRNAs (hsa-let-7a-3p, hsa-let-7g-5p, has-miR-141-3p, and has-miR-29b-3p). We can verify that regulation of CCND2 gene by let-7a miRNA plays an important role in cancer development from a literature [113].

In summary, we can verify that miRTarVis is an effective tool for analysis of miRNA-mRNA expression data by applying miRTarVis to TCGA breast cancer data. The visualizations generated by miRTarVis effectively help a user to check previously reported miRNA-mRNA interactions from literatures. This shows the possibility that the predicted miRNA-mRNA interactions by miRTarVis may include not-yet-verified set of miRNA-mRNA interactions that can be verified by a well-design biological experiment. Therefore, use of miRTarVis can

help biologists to make a hypothesis by exploring their miRNA-mRNA expression data with various options.

Table 5.3. This table shows the 200 predicted miRNA-mRNA interactions from TCGA breast cancer data by miRTarVis in our case study.

| miRNA | mRNA | Correlation | MI | GenMiR++ | MINE | TargetScan | miRanda |
|---|---|---|---|---|---|---|---|
| hsa-miR-96-5p | PIK3R1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-96-5p | TRIM9 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-96-5p | MAK | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-96-5p | CABP1 | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-96-5p | PTPRE | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-96-5p | SPRED1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-96-5p | NTN4 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-96-5p | NCALD | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-96-5p | NLGN4X | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-214-3p | INCENP | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-214-3p | CELF1 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-301a-3p | PRNP | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-195-3p | PHKA1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-195-3p | RAP2C | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-195-3p | PAK2 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | CNTFR | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | EPHA4 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | GNG12 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-21-5p | CHL1 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | RGS7BP | TRUE | FALSE | FALSE | TRUE | TRUE | FALSE |
| hsa-miR-21-5p | MATN2 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | SNX29 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | FAM13A | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-21-5p | FGF1 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | VCL | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | ARHGAP31 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | THRB | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | BNC2 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-21-5p | ZBTB47 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-21-5p | KLF3 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-21-5p | RECK | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-let-7g-5p | SLC46A3 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | IKZF2 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | RPS6KA3 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | ARRDC4 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | EPHA4 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | HOXD1 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | FIGNL2 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | INPP5A | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | DDX26B | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | IGF1 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | STAB2 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | C20orf194 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | RNF165 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | AKAP6 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | MGAT3 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | DMD | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | APBA1 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | LY75 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | C3orf64 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | CYP46A1 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | TMEM143 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | CCND2 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | RIMS3 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | RGS6 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | SDK1 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | COL14A1 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | ADAMTS5 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | ADAMTS1 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7g-5p | HLF | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | LRIG3 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | CMTM6 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7g-5p | MEIS2 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-miR-30e-3p | CHORDC1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-30e-3p | SFXN1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-30e-3p | CELF1 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-30e-3p | XPO1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-30e-3p | RHEBL1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |

| miRNA | mRNA | Correlation | MI | GenMiR++ | MINE | TargetScan | miRanda |
|---|---|---|---|---|---|---|---|
| hsa-miR-27a-5p | GRSF1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-27a-5p | MESDC1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-27a-5p | CDK5R1 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | DCAF5 | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-29b-3p | VPS36 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | TTYH2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | ADAMTS18 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | PCDHA5 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | HBP1 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | TRIM9 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | C11orf87 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | KCNMA1 | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-29b-3p | RARB | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | DAB2IP | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-29b-3p | NUDT11 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | NDN | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | LRCH1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | ADAMTS9 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | NLGN3 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-29b-3p | COL15A1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | WASF2 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-29b-3p | TSPAN14 | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-29b-3p | ARRDC4 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | ISLR2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | MAP6 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | N4BP2L1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | MAPRE2 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | SVIL | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | CCND2 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | SIDT2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | SDK1 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | DLG2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | ARHGEF10 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | MMP2 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-29b-3p | HLF | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | FAM168B | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | NAP1L3 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-29b-3p | FAM167A | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-26a-5p | CHORDC1 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-10b-3p | SSX2IP | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-15b-5p | MAK | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | SLC46A3 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7a-3p | IKZF2 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7a-3p | ARRDC4 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | EPHA4 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | HOXD1 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | FIGNL2 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | STAB2 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7a-3p | KIAA1274 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | RNF165 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | MGAT3 | TRUE | TRUE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7a-3p | APBA1 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7a-3p | CYP46A1 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | TMEM143 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | CCND2 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7a-3p | RGS6 | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | CHRD | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | SDK1 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE |
| hsa-let-7a-3p | COL14A1 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | SCN4B | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | HLF | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | LRIG3 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | CMTM6 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7a-3p | MEIS2 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-let-7c-5p | TMPO | TRUE | FALSE | TRUE | TRUE | TRUE | FALSE |
| hsa-let-7c-5p | TMEM194A | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE |

| miRNA | mRNA | Correlation | MI | GenMiR++ | MINE | TargetScan | miRanda |
|---|---|---|---|---|---|---|---|
| hsa-let-7c-5p | PANX2 | TRUE | FALSE | TRUE | TRUE | TRUE | FALSE |
| hsa-let-7c-5p | MESDC1 | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE |
| hsa-let-7c-5p | ZNF200 | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE |
| hsa-miR-320a | HNRNPC | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-320a | ARF1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-320a | ACTL6A | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-320a | YTHDF1 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-320a | RGS9BP | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-320a | HSPA4 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-320a | CKAP5 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-320a | E2F1 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-141-3p | TTYH2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | CDC14A | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | MAF | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-141-3p | FRMD6 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | IKZF2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | CCNL2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | STAT5B | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | IGF2 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-141-3p | CCND2 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | WNT5A | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | SIDT2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | SOBP | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | LRRTM2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-141-3p | HLF | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-125b-5p | ANKRD13B | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-125b-5p | POU2F1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-125b-5p | CPSF6 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-125b-5p | DNAJC14 | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-125b-5p | GRSF1 | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-125b-5p | ZDHHC9 | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-485-5p | SRGAP2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-326 | GRPEL2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-326 | RBBP4 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-15a-5p | DLST | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-15a-5p | MAK | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-15a-5p | MOBKL2B | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-497-5p | PHKA1 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-497-5p | RAP2C | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-195-5p | FAM122B | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-195-5p | PHKA1 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-195-5p | TMEM33 | TRUE | FALSE | FALSE | TRUE | TRUE | FALSE |
| hsa-miR-21-3p | RPS6KA3 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-340-3p | INPP5A | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-340-3p | LPHN2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-182-5p | PIK3R1 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-182-5p | SPRED1 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-1271-5p | MTCH2 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-1271-5p | WHSC1 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-148b-5p | PRNP | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-148b-5p | SH3TC2 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-454-3p | FNDC4 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-10b-5p | SSX2IP | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-10b-5p | E2F7 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-22-3p | PFDN6 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-15b-3p | CDC42EP2 | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-15b-3p | MAK | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-15b-3p | LPHN2 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-378c | PSMA5 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-378c | XPO5 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-378c | GRSF1 | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE |
| hsa-miR-143-3p | DLG3 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-143-3p | ENSA | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-143-3p | WHSC1 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE |
| hsa-miR-183-5p | SIRPA | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |
| hsa-miR-33a-5p | FAM129A | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE |

# Chapter 6

# Discussion

In this paper, we presented a visual analysis tool for miRNA-mRNA expression data. MiRNA is an important factor for gene regulation. Many bioinformatician have interest on miRNA. We concentrated the expression value of miRNA to analyze the regulation effect of MiRNA to mRNA expression data. We also used the sequence information of miRNA to effect mRNA expression by adopting sequence based miRNA prediction from Bioinformatics field.

There can be other attributes of miRNAs that can be considered in miRNA analysis. For example, the genomic location of miRNA can effect gene regulation of miRNA. Some miRNA genes are from the intron region (non-protein coding sub-region of a protein coding gene) of a gene. MiRNAs from the intron of a gene may have more regulatory relationship

with that gene. However, in current design of miRTarVis, the genomic location of miRNA is not considered in the analysis and visualization procedure. As a future work, we will devise a new visual analysis technique and analysis technique that can represent the genomic location of a miRNA and its effectiveness over gene regulation.

There are other gene regulators for the study of epigenomics. For example, DNA methylation is known to have great relationship with gene regulation in animal cells. Histone modification is another gene regulation factor that researchers have great interest. Transcription factor (TF) genes are a gene that regulates another gene' s expression. In this paper, the scope of our discussion is about miRNAs that are predicted to regulate the gene expression.

If we could expand the scope to DNA methylation, histone modification, and regulation factor gene, it would be very helpful for those researchers who want to analyze various types of epigenomics data. The integrated analysis can widen the analysis capability of the researchers, and can suggest a meaningful new research hypothesis. If we have the chance to update and upgrade the miRTarVis for next

version of it, we want to consider about the integration of various epigenomics data for visualization all the epigenomics data simultaneously.

miRTarVis handles expression profile data. In genomics and epigenomics expeirments, expression profile data were usually measure by Microarray methods. However, as next generation sequencing techniques emerges, sequencing data is generated for measureing miRNA and mRNA expression profile data. In miRTarVis, the input if the processed expression level data. However, in sequencing experiments, the raw data is not matrix-like expression profile table, but many short sequences from a target cell.

In terms of Information Visualization we presented a new visual encoding technique for bipartite network data, suggesting Bipartite Treemap. Bipartite Treemap can be a good solution when the target network suffers occlusion problem among nodes or edges in node-link diagram by the node-link diagram. Moreover, Bipartite Treemap efficiently encodes the edges' attributes by the size of the rectangle. Therefore, Bipartite Treemap can be useful not only for miRNA-mRNA

prediction network, but for all bipartite networks. The data such as student-teacher relationship in a school may be a possible bipartite network that can be visualized by Bipartite Treemap technique. As a future work, we want to apply Bipartite Treemap to another bipartite network from other research fields such as social network data analysis.

Our Bipartite Treemap for miRNA-mRNA regulatory network sets miRNA as top level factor. We designed like that because of the characteristic of predicted miRNA-mRNA network. The number of miRNAs is much smaller than the number of mRNAs, and the number of linked mRNAs to one miRNA is much greater than the number of linked miRNAs to one mRNA. However, there is a demand to see mRNA as top level. If Bipartite Treemap can change the hierarchy according to users' demand, it can help users to explore the network more effectively. For example, if miRNA-mRNA regulatory network is visualized as mRNA being the top level and miRNA being the down level, it is more easier to see what miRNA are predicted to regulate a specific mRNA. Therefore, we want to enable users to convert the hierarchy easily in Bipartite Treemap in future work.

In our enhanced node-link diagram, we added many visualization improvements and interactions for effective analysis for miRNA-mRNA interactions from network by miRNA prediction algorithms. We deliberated over what informations users want to see directly in the visualization. Basically, the most important information that users want to see is the miRNA-mRNA target interaction. In previous work, it is difficult to comprehend the network structure explicitly when the network is congested by many miRNA-mRNA interactions. To show the edges clearer, our node-link diagram's edges have alpha value, in other words, they are transparent. This transparency of edges helps users see links with less occlusion problem.

# Chapter 7

# Conclusion

In this dissertation, we presented Bipartite Treemap and enhanced node-link diagram as a new visualization technique for miRNA-mRNA interaction network and introduced miRTarVis, an interactive visual analysis tool for miRNA-mRNA expression profile data.

Bipartite Treemap exploits the properties of miRNA-mRNA interaction network for more effective visualization of miRNA-mRNA interaction network. Enhanced node-link diagram shows overall structure of the network well, and we added more user interaction for users to easily explore their miRNA-mRNA expression profile data with the node-link diagram.

For designing miRTarVis, we defined a representative analysis pipeline for the data and designed mirTarVis to support the analysis

pipeline based on a foldable accordion metaphor. miRTarVis integrates various miRNA target prediction algorithms, including a novel prediction algorithm using the MINE analysis, into the analysis pipeline. miRTarVis shows the resulting miRNA target network using interactive Bipartite Treemap and enhanced node-link diagram.

We conducted a case study to prove the efficacy of miRTarVis. We analyzed a miRNA-mRNA expression profile data of asthma patients and found a potentially novel mechanism by which adiposity increases fibrosis in asthma. We conducted another case study with TCGA breast cancer miRNA-mRNA expression profile data, and found important miRNAs and mRNAs that are known to play an important role in cancer development.

In summary, this work's contributions are three-fold: (1) introducing new visualization techniques for more effective visualization of predicted miRNA-mRNA interaction network, (2) seamlessly integrating sequence-based prediction algorithms and miRNA-mRNA expression profile based prediction algorithms, and (3) presenting a visual analysis tool for miRNA-mRNA expression profile

data for more easy data exploration for miRNA-mRNA expression profile data study.

# Bibliography

[1] J. P. de Magalhães, C. E. Finch and G. Janssens, "Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions," *Ageing research reviews,* vol. 9, no. 3, pp. 315-323, 2010.

[2] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo and X. Liu, "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells," *PLoS ONE,* vol. 9, no. 1, pp. 1-13, 2014.

[3] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nat Rev Genet,* vol. 10, no. 10, pp. 669-680, 2009.

[4] M. Baker, "Next-generation sequencing: adjusting to data overload," *nature methods,* vol. 7, no. 7, pp. 495-499, 2010.

[5] S. E. Baranzini, J. Mudge, J. C. van Velkinburgh, P. Khankhanian, I. Khrebtukova, N. A. Miller, L. Zhang, A. D. Farmer, C. J. Bell, R. W. Kim and others, "Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis," *Nature,* vol. 464, no. 7293, pp. 1351-1356, 2010.

[6] S. Zhang, Q. Li, J. Liu and X. J. Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules," *Bioinformatics,* vol. 27, no. 13, pp. i401--i409, 2011.

[7] J. Zhang, R. Chiodini, A. Badr and G. Zhang, "The impact of next-generation sequencing on genomics," *Journal of genetics and genomics,* vol. 38, no. 3, pp. 95-109, 2011.

[8] J. A. Martin and Z. Wang, "Next-generation transcriptome assembly," *Nature Reviews Genetics,* vol. 12, no. 10, pp. 671-682, 2011.

[9] A. Teufel, M. Krupp, A. Weinmann and P. R. Galle, "Current bioinformatics tools in genomic biomedical research (Review)," *International journal of molecular medicine,* vol. 17, no. 6, pp. 967-973, 2006.

[10] R. Jaenisch and A. Bird, "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals," *Nature genetics,* vol. 33, pp. 245-254, 2003.

[11] K. V. Morris and J. S. Mattick, "The rise of regulatory RNA," *Nature reviews. Genetics,* vol. 15, no. 6, p. 423, 2014.

[12] D. P. Bartel, "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function," *Cell ,* vol. 116, no. 2, pp. 281-297, 2004.

[13] T. M. Williams, J. E. Selegue, T. Werner, N. Gompel, A. Kopp and S. B. Carroll, "The regulation and evolution of a genetic switch controlling sexually dimorphic traits in Drosophila," *Cell,* vol. 134, no. 4, pp. 610-623, 2008.

[14] A. Muniategui, J. Pey, F. J. Planes and A. Rubio, "Joint analysis of miRNA and mRNA expression data," *Briefings in Bioinformatics,* vol. 14, no. 3, pp. 263-278, 2013.

[15] H. Jin, W. Tuo, H. Lian, Q. Liu, X.-Q. Zhu and H. Gao, "Strategies to identify microRNA targets: new advances," *New biotechnology,* vol. 27, no. 6, pp. 734-738, 2010.

[16] D. W. Thomson, C. P. Bracken and G. J. Goodall, "Experimental strategies for microRNA target identification," *Nucleic acids research,* vol. 39, no. 16, pp. 6845-6853, 2011.

[17] T. M. Witkos, E. Koscianska and W. J. Krzyzosiak, "Practical aspects of microRNA target prediction," *Current molecular medicine,* vol. 11, no. 2, pp. 93-109, 2011.

[18] B. P. Lewis, C. B. Burge and D. P. Bartel, "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets," *Cell ,* vol. 120, no. 1, pp. 15-20, 2005.

[19] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, D. S. Marks and others, "MicroRNA targets in Drosophila," *Genome biology,* vol. 5, no. 1, pp. R1--R1, 2004.

[20] J. C. Huang, T. Babak, T. W. Corson, G. Chua, S. Khan, B. L. Gallie, T. R. Hughes, B. J. Blencowe, B. J. Frey and Q. D. Morris, "Using expression profiling data to identify human microRNA targets," *Nature methods,* vol. 4, no. 12, pp. 1045-1049, 2007.

[21] J. Li, R. Min, A. Bonner and Z. Zhang, "A probabilistic framework to improve microrna target prediction by incorporating proteomics data," *Journal of bioinformatics and computational biology,* vol. 7, no. 06, pp. 955-972, 2009.

[22] K.-O. Mutz, A. Heilkenbrinker, M. Lönne, J.-G. Walter and F. Stahl, "Transcriptome analysis using next-generation sequencing," *Current Opinion in Biotechnology ,* vol. 24, no. 1, pp. 22-30, 2013.

[23] K. J. Mantione, R. M. Kream, H. Kuzelova, R. Ptacek, J. Raboch, J. M. Samuel and G. B. Stefano, "Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq," *Medical Science Monitor Basic Research,* vol. 20, pp. 138-141, #jul# 2014.

[24] A. Git, H. Dvinge, M. Salmon-Divon, M. Osborne, C. Kutter, J. Hadfield, P. Bertone and C. Caldas, "Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression," *Rna,* vol. 16, no. 5, pp. 991-1006, 2010.

[25] C. C. Pritchard, H. H. Cheng and M. Tewari, "MicroRNA profiling: approaches and considerations," *Nature Reviews Genetics,* vol. 13, no. 5, pp. 358-369, 2012.

[26] H. Willenbrock, J. Salomon, R. Søkilde, K. B. Barken, T. N. Hansen, F. C. Nielsen, S. Møller and T. Litman, "Quantitative miRNA expression analysis: Comparing microarrays with next-generation sequencing," *RNA,* vol. 15, no. 11, pp. 2028-2034, 2009.

[27] S. Nam, M. Li, K. Choi, C. Balch, S. Kim and K. P. Nephew, "MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression," *Nucleic acids research,* p. gkp294, 2009.

[28] G. T. Huang, C. Athanassiou and P. V. Benos, "mirConnX: condition-specific mRNA-microRNA network integrator," *Nucleic acids research,* p. gkr276, 2011.

[29] G. Sales, A. Coppe, A. Bisognin, M. Biasiolo, S. Bortoluzzi and C. Romualdi, "MAGIA, a web-based tool for miRNA and Genes Integrated Analysis," *Nucleic acids research,* p. gkq423, 2010.

[30] A. Bisognin, G. Sales, A. Coppe, S. Bortoluzzi and C. Romualdi, "MAGIA2: from miRNA and genes expression data integrative analysis to microRNA--transcription factor mixed regulatory circuits (2012 update)," *Nucleic acids research,* p. gks460, 2012.

[31] D. A. Keim, "Information visualization and visual data mining," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 8, no. 1, pp. 1-8, 2002.

[32] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*, 1996.

[33] H. Chae, S. Rhee, K. P. Nephew and S. Kim, "BioVLAB-MMIA-NGS: microRNA–mRNA integrated analysis using high-throughput sequencing data," *Bioinformatics,* vol. 31, no. 2, pp. 265-267, 2015.

[34] S. Cho, I. Jang, Y. Jun, S. Yoon, M. Ko, Y. Kwon, I. Choi, H. Chang, D. Ryu, B. Lee, V. N. Kim, W. Kim and S. Lee, "miRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting," *Nucleic Acids Research,* vol. 41, no. D1, pp. D252-D257, 2013.

[35] B. Shneiderman and M. Wattenberg, "Ordered treemap layouts," in *infovis*, 2001.

[36] E. Wienholds and R. H. A. Plasterk, "MicroRNA function in animal development," *FEBS letters,* vol. 579, no. 26, pp. 5911-5922, 2005.

[37] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang and Y. Liu, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic acids research,* vol. 37, no. suppl 1, pp. D98--D104, 2009.

[38] S. M. Hammond, "An overview of microRNAs," *Advanced drug delivery reviews,* vol. 87, pp. 3-14, 2015.

[39] N. Rajewsky, "microRNA target predictions in animals," *Nature genetics,* vol. 38, pp. S8--S13, 2006.

[40] X. Fan and L. Kurgan, "Comprehensive overview and assessment of computational prediction of microRNA targets in animals," *Briefings in Bioinformatics,* vol. 16, no. 5, pp. 780-794, 2015.

[41] A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel and N. Rajewsky, "Combinatorial microRNA target predictions," *Nat Genet,* vol. 37, no. 5, pp. 495-500, #may# 2005.

[42] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic acids research,* vol. 42, no. D1, pp. D68--D73, 2014.

[43] J.-M. Claverie, "What if there are only 30,000 human genes?," *Science,* vol. 291, no. 5507, pp. 1255-1257, 2001.

[44] J.-i. Takeda, Y. Suzuki, R. Sakate, Y. Sato, T. Gojobori, T. Imanishi and S. Sugano, "H-DBAS: human-transcriptome database for alternative splicing: update 2010," *Nucleic acids research,* vol. 38, no. suppl 1, pp. D86--D90, 2010.

[45] W. Filipowicz, S. N. Bhattacharyya and N. Sonenberg, "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?," *Nature Reviews Genetics,* vol. 9, no. 2, pp. 102-114, 2008.

[46] M. E. Peter, "Targeting of mRNAs by multiple miRNAs: the next step," *Oncogene,* vol. 29, no. 15, pp. 2161-2164, 2010.

[47] N. K. Vo, R. P. Dalton, N. Liu, E. N. Olson and R. H. Goodman, "Affinity purification of microRNA-133a with the cardiac transcription factor, Hand2," *Proceedings of the National Academy of Sciences,* vol. 107, no. 45, pp. 19231-19236, 2010.

[48] M. Thomas, J. Lieberman and A. Lal, "Desperately seeking microRNA targets," *Nature structural \& molecular biology,* vol. 17, no. 10, pp. 1169-1174, 2010.

[49] K. C. Miranda, T. Huynh, Y. Tay, Y.-S. Ang, W.-L. Tam, A. M. Thomson, B. Lim and I. Rigoutsos, "A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes," *Cell,* vol. 126, no. 6, pp. 1203-1217, 2006.

[50] S. Wu, S. Huang, J. Ding, Y. Zhao, L. Liang, T. Liu, R. Zhan and X. He, "Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3 untranslated region," *Oncogene,* vol. 29, no. 15, pp. 2302-2308, 2010.

[51] I. Lee, S. S. Ajay, J. I. Yook, H. S. Kim, S. H. Hong, N. H. Kim, S. M. Dhanasekaran, A. M. Chinnaiyan and B. D. Athey, "New class of microRNA targets containing simultaneous 5 -UTR and 3 -UTR interaction sites," *Genome research,* vol. 19, no. 7, pp. 1175-1183, 2009.

[52] A. Brümmer and J. Hausser, "MicroRNA binding sites in the coding region of mRNAs: Extending the repertoire of post-transcriptional gene regulation," *Bioessays,* vol. 36, no. 6, pp. 617-626, 2014.

[53] A. E. Pasquinelli, "MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship," *Nature Reviews Genetics,* vol. 13, no. 4, pp. 271-282, 2012.

[54] D. A. Keim, F. Mansmann, J. Schneidewind and H. Ziegler, "Challenges in visual data analysis," in *Information Visualization, 2006. IV 2006. Tenth International Conference on*, 2006.

[55] H. Zheng, R. Fu, J.-T. Wang, Q. Liu, H. Chen and S.-W. Jiang, "Advances in the Techniques for the Prediction of microRNA Targets," *International journal of molecular sciences,* vol. 14, no. 4, pp. 8179-8187, 2013.

[56] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell,* vol. 136, no. 2, pp. 215-233, 2009.

[57] T. Saito and P. Sætrom, "MicroRNAs – targeting and target prediction," *New Biotechnology ,* vol. 27, no. 3, pp. 243-249, 2010.

[58] P. Alexiou, M. Maragkakis, G. L. Papadopoulos, M. Reczko and A. G. Hatzigeorgiou, "Lost in translation: an assessment and perspective for computational microRNA target identification," *Bioinformatics,* vol. 25, no. 23, pp. 3049-3055, 2009.

[59] D. Betel, M. Wilson, A. Gabow, D. S. Marks and C. Sander, "The microRNA. org resource: targets and expression," *Nucleic acids research,* vol. 36, no. suppl 1, pp. D149--D153, 2008.

[60] D. Betel, A. Koppal, P. Agius, C. Sander and C. Leslie, "Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites," *Genome biology,* vol. 11, no. 8, p. R90, 2010.

[61] C.-H. Chou, N.-W. Chang, S. Shrestha, S.-D. Hsu, Y.-L. Lin, W.-H. Lee, C.-D. Yang, H.-C. Hong, T.-Y. Wei, S.-J. Tu, T.-R. Tsai, S.-Y. Ho, T.-Y. Jian, H.-Y. Wu, P.-R. Chen, N.-C. Lin, H.-T. Huang, T.-L. Yang, C.-Y. Pai, C.-S. Tai, W.-L. Chen, C.-Y. Huang, C.-C. Liu, S.-L. Weng, K.-W. Liao, W.-L. Hsu and H.-D. Huang, "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database," *Nucleic Acids Research,* vol. 44, no. D1, pp. D239-D247, 2016.

[62] V. Agarwal, G. W. Bell, J.-W. Nam and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *Elife,* vol. 4, p. e05005, 2015.

[63] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis and others, "DIANA-microT web server: elucidating microRNA functions through target prediction," *Nucleic acids research,* p. gkp292, 2009.

[64] S. Griffiths-Jones, R. J. Grocock, S. Van Dongen, A. Bateman and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic acids research,* vol. 34, no. suppl 1, pp. D140--D144, 2006.

[65] M. M. Shareef, B. Isaac, S. Philip, T. Udayakumar and A. Pollack, "AKT in Differential miRNA Processing in Prostate Carcinoma," 2013.

[66] J. Fu, W. Tang, P. Du, G. Wang, W. Chen, J. Li, Y. Zhu, J. Gao and L. Cui, "Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis," *BMC systems biology,* vol. 6, no. 1, p. 68, 2012.

[67] F. Allantaz, D. T. Cheng, T. Bergauer, P. Ravindran, M. F. Rossier, M. Ebeling, L. Badi, B. Reis, H. Bitter, M. D'Asaro and others, "Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression," *PloS one,* vol. 7, no. 1, p. e29979, 2012.

[68] C. Camps, H. K. Saini, D. R. Mole, H. Choudhry, M. Reczko, J. A. Guerra-Assunção, Y.-M. Tian, F. M. Buffa, A. L. Harris, A. G. Hatzigeorgiou and others, "Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia," *Molecular cancer,* vol. 13, no. 1, p. 1, 2014.

[69] L. M. Simon, L. C. Edelstein, S. Nagalla, A. B. Woodley, E. S. Chen, X. Kong, L. Ma, P. Fortina, S. Kunapuli, M. Holinstat and others, "Human platelet microRNA-mRNA networks associated with age and gender revealed by integrated plateletomics," *Blood,* vol. 123, no. 16, pp. e37--e45, 2014.

[70] L. Song, P. Langfelder and S. Horvath, "Comparison of co-expression measures: mutual information, correlation, and model based indices," *BMC bioinformatics,* vol. 13, no. 1, p. 328, 2012.

[71] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, "Detecting novel associations in large data sets," *science,* vol. 334, no. 6062, pp. 1518-1524, 2011.

[72] D. Jung, B. Kim, R. J. Freishtat, M. Giri, E. Hoffman and J. Seo, "miRTarVis: an interactive visual analysis tool for microRNA-mRNA expression profile data," in *BMC proceedings*, 2015.

[73] A. A. Khan, D. Betel, M. L. Miller, C. Sander, C. S. Leslie and D. S. Marks, "Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs," *Nature biotechnology,* vol. 27, no. 6, pp. 549-555, 2009.

[74] D. M. Garcia, D. Baek, C. Shin, G. W. Bell, A. Grimson and D. P. Bartel, "Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs," *Nature structural \& molecular biology,* vol. 18, no. 10, pp. 1139-1146, 2011.

[75] X. Li, R. Gill, N. G. F. Cooper, J. K. Yoo and S. Datta, "Modeling microRNA-mRNA interactions using PLS regression in human colon cancer," *BMC medical genomics,* vol. 4, no. 1, p. 1, 2011.

[76] A. Muniategui, R. Nogales-Cadenas, M. Vázquez, X. L. Aranguren, X. Agirre, A. Luttun, F. Prosper, A. Pascual-Montano and A. Rubio, "Quantification of miRNA-mRNA Interactions," *PLoS ONE,* vol. 7, no. 2, p. e30766, #feb# 2012.

[77] F. C. Stingo, Y. A. Chen, M. Vannucci, M. Barrier and P. E. Mirkes, "A Bayesian graphical modeling approach to microRNA regulatory network inference," *The annals of applied statistics,* vol. 4, no. 4, p. 2024, 2010.

[78] V. Fulci, T. Colombo, S. Chiaretti, M. Messina, F. Citarella, S. Tavolaro, A. Guarini, R. Foà and G. Macino, "Characterization of B- and T-lineage acute lymphoblastic leukemia by integrated analysis of MicroRNA and mRNA expression profiles," *Genes, Chromosomes and Cancer,* vol. 48, no. 12, pp. 1069-1082, 2009.

[79] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub, "MicroRNA expression profiles classify human cancers," *Nature,* vol. 435, no. 7043, pp. 834-838, 2005.

[80] M. Lionetti, M. Biasiolo, L. Agnelli, K. Todoerti, L. Mosca, S. Fabris, G. Sales, G. L. Deliliers, S. Bicciato, L. Lombardi, S. Bortoluzzi and A. Neri, "Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma," *Blood,* vol. 114, no. 25, pp. e20--e26, 2009.

[81] N. C. Gutierrez, M. E. Sarasquete, I. Misiewicz-Krzeminska, M. Delgado, J. De Las Rivas, F. V. Ticona, E. Ferminan, P. Martin-Jimenez, C. Chillon, A. Risueno, J. M. Hernandez, R. Garcia-Sanz, M. Gonzalez and J. F. San Miguel, "Deregulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling," *Leukemia,* vol. 24, no. 3, pp. 629-637, #mar# 2010.

[82] Y.-P. Wang and K.-B. Li, "Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data," *BMC genomics,* vol. 10, no. 1, p. 218, 2009.

[83] T. D. Le, J. Zhang, L. Liu and J. Li, "Ensemble Methods for MiRNA Target Prediction from Expression Data," *PloS one,* vol. 10, no. 6, p. e0131627, 2015.

[84] T. D. Le, L. Liu, A. Tsykin, G. J. Goodall, B. Liu, B.-Y. Sun and J. Li, "Inferring microRNA–mRNA causal regulatory relationships from expression data," *Bioinformatics,* vol. 29, no. 6, pp. 765-771, 2013.

[85] R. J. a. M. D. a. S.-R. J. a. S. P. K. a. A. L. G. a. X. X. a. C. N. D. a. A.-B. G. a. S. G. Prill, "Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges," *PLoS ONE,* vol. 5, no. 2, pp. 1-18, 02 2010.

[86] Y. Li, C. Liang, K.-C. Wong, K. Jin and Z. Zhang, "Inferring probabilistic miRNA–mRNA interaction signatures in cancers: a role-switch approach," *Nucleic Acids Research,* vol. 42, no. 9, p. e76, 2014.

[87] R. Søkilde, B. Kaczkowski, A. Podolska, S. Cirera, J. Gorodkin, S. Møller and T. Litman, "Global microRNA Analysis of the NCI-60 Cancer Cell Panel," *Molecular Cancer Therapeutics,* vol. 10, no. 3, pp. 375-384, 2011.

[88] M. Riaz, M. T. M. van Jaarsveld, A. Hollestelle, W. J. C. Prager-van der Smissen, A. A. J. Heine, A. W. M. Boersma, J. Liu, J. Helmijr, B. Ozturk, M. Smid, E. A. Wiemer, J. A. Foekens and J. W. M. Martens, "miRNA expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific miRNAs," *Breast Cancer Research,* vol. 15, no. 2, pp. 1-17, 2013.

[89] J. Qin, M. J. Li, P. Wang, N. S. Wong, M. P. Wong, Z. Xia, G. S. W. Tsao, M. Q. Zhang and J. Wang, "ProteoMirExpress: Inferring MicroRNA and Protein-centered Regulatory Networks from High-throughput Proteomic and mRNA Expression Data," *Molecular & Cellular Proteomics,* vol. 12, no. 11, pp. 3379-3387, 2013.

[90] G. Pio, M. Ceci, D. Malerba and D. D'Elia, "ComiRNet: a web-based system for the analysis of miRNA-gene regulatory networks," *BMC bioinformatics,* vol. 16, no. Suppl 9, p. S7, 2015.

[91] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul and E. Segal, "The role of site accessibility in microRNA target recognition," *Nat Genet,* vol. 39, no. 10, pp. 1278-1284, #oct# 2007.

[92] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *STAT. APPL. GENET. MOL. BIOL,* vol. 3, no. 1, 2004.

[93] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome biol,* vol. 11, no. 10, p. R106, 2010.

[94] L. Goff, C. Trapnell and D. Kelley, "cummeRbund: analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data," *R package version,* vol. 2, no. 0, 2012.

[95] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, Springer, 2005, pp. 397-420.

[96] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research,* vol. 13, no. 11, pp. 2498-2504, 2003.

[97] M. Bruls, K. Huizing and J. J. Van Wijk, Squarified treemaps, Springer, 2000.

[98] B. Meyer, "Self-organizing graphs—a neural network perspective of graph layout," in *Graph Drawing*, 1998.

[99] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Information processing letters,* vol. 31, no. 1, pp. 7-15, 1989.

[100] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. I. Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug and others, "GeneCards Version 3: the human gene integrator," *Database,* vol. 2010, p. baq020, 2010.

[101] L. Hiddingh, R. S. Raktoe, J. Jeuken, E. Hulleman, D. P. Noske, G. J. L. Kaspers, W. P. Vandertop, P. Wesseling and T. Wurdinger, "Identification of temozolomide resistance factors in glioblastoma via integrative miRNA/mRNA regulatory network analysis," *Scientific reports,* vol. 4, 2014.

[102] M. D. Robinson, D. J. McCarthy and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, no. 1, pp. 139-140, 2010.

[103] S. Griffiths-Jones, H. K. Saini, S. van Dongen and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic acids research,* vol. 36, no. suppl 1, pp. D154--D158, 2008.

[104] G. Anders, S. D. Mackowiak, M. Jens, J. Maaskola, A. Kuntzagk, N. Rajewsky, M. Landthaler and C. Dieterich, "doRiNA: a database of RNA interactions in post-transcriptional regulation," *Nucleic acids research,* p. gkr1007, 2011.

[105] X. Wang, "miRDB: a microRNA target prediction and functional annotation database with a wiki interface," *Rna,* vol. 14, no. 6, pp. 1012-1017, 2008.

[106] Y. Ru, K. J. Kechris, B. Tabakoff, P. Hoffman, R. A. Radcliffe, R. Bowler, S. Mahaffey, S. Rossi, G. A. Calin, L. Bemis and D. Theodorescu, "The multiMiR R package and database: integration of microRNA–target interactions along with their disease and drug associations," *Nucleic Acids Research,* vol. 42, no. 17, p. e133, 2014.

[107] A. Kraskov, H. Stögbauer and P. Grassberger, "Estimating mutual information," *Phys. Rev. E,* vol. 69, p. 066138, Jun 2004.

[108] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman and C. Furlanello, "Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers," *Bioinformatics,* p. bts707, 2012.

[109] S. C. Ferrante, E. P. Nadler, D. K. Pillai, M. J. Hubal, Z. Wang, J. M. Wang, H. Gordish-Dressman, E. Koeck, S. Sevilla, A. A. Wiles and others, "Adipocyte-derived exosomal miRNAs: a novel mechanism for obesity-related disease," *Pediatric research,* vol. 77, no. 3, pp. 447-454, 2014.

[110] Y. Pekarsky, U. Santanam, A. Cimmino, A. Palamarchuk, A. Efanov, V. Maximov, S. Volinia, H. Alder, C.-G. Liu, L. Rassenti and others, "Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181," *Cancer research,* vol. 66, no. 24, pp. 11590-11593, 2006.

[111] J. L. Mott, S. Kobayashi, S. F. Bronk and G. J. Gores, "mir-29 regulates Mcl-1 protein expression and apoptosis," *Oncogene,* vol. 26, no. 42, pp. 6133-6140, 2007.

[112] S.-Y. Park, J. H. Lee, M. Ha, J.-W. Nam and V. N. Kim, "miR-29 miRNAs activate p53 by targeting p85 and CDC42," *Nature structural molecular biology,* vol. 16, no. 1, pp. 23-29, 2009.

[113] Q. Dong, P. Meng, T. Wang, W. Qin, W. Qin, F. Wang, J. Yuan, Z. Chen, A. Yang and H. Wang, "MicroRNA let-7a inhibits proliferation of human prostate cancer cells in vitro and in vivo by targeting E2F2 and CCND2," *PloS one,* vol. 5, no. 4, p. e10147, 2010.

# 요약

본 연구에서는 마이크로 알엔에이-엠알엔에이 표현형 프로파일 데이터 분석을 위한 시각적 분석 기법을 제시한다. 마이크로알엔에이는 그것의 타겟 유전자의 발현을 억제하는 짧은 길이의 뉴클레오 타이드로서, 많이 마이크로알엔에이 타겟 예측 알고리즘들이 마이크로 알엔에이와 그것의 타겟의 염기서열 정보를 활용하였다. 최근에는 마이크로알엔에이-엠알엔에이 표현형 프로파일 데이터를 활용한 타겟 예측이 제시되었다. 기존의 웹 기반의 분석 도구들이 마이크로알엔에이-엠알엔에이 익스프레션 프로파일 데이터로부터 타겟을 예측하기 위해 제시되었지만, 분석 능력이 부족할 뿐 아니라, 인터랙티브한 시각화를 통한 정보 제시가 미비한 실정이다. 본 논문에서는 바이퍼타이트 트리맵 (Bipartite Treemap) 과 개선된 노드링크 다이어그램 (Enhanced Node-Link Diagram) 이라는 두 개의 시각화 기법을 제시하여, 사용자가 효율적으로 마이크로알엔에이-엠알엔에이 익스프레션 프로파일 데이터를 시각적으로 분석할 수 있도록 하였다. 또한 이 두 가지 기법을 바탕으로 멀타비즈 (miRTarVis) 라는 마이크로알엔에이-엠알엔에이 익스프레션 프로파일 데이터 분석 도구를 제시하였다. 새로운 시각적 분석 기법과 분석 도구로서의 멀타비즈의 유용성을 검증하기 위해, 천식-비천식 환자에 대한 실험에서 비롯된 마이크로알엔에이-엠알엔에이 표현형 데이터에 멀타비즈를 적용하였다. 그 결과, 마이크로알엔에이가 유전자 발현에 미치는 영향을 쉽게 확인할 수 있음을 확인하였다. 또한 기존의 다른 시각화 도구와 멀타비즈를 비교하여, 멀타비즈가 가지는 유용성을 객관적으로 검증하였다.

**주요어:** 마이크로알엔에이 (MicroRNA), 엠알엔에이 (mRNA), 시각화, 유전자 표현형, 마이크로알엔에이 타겟 예측, 시각적 분석, 바이퍼타이트 트리맵 (Bipartite Treemap), 멀타비즈 (miRTarVis)

**학번:** 2010-20886